# Transcriptional and Epigenetic Dynamics Observed During Lineage Specification of Human Embryonic Stem Cells

# Transcriptional and Epigenetic Dynamics Observed During Lineage Specification of Human Embryonic Stem Cells

A dissertation presented

by

**Casey Gifford**

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

September, 2013

Dissertation advisor: Dr. Alexander Meissner Casey Gifford

**Transcriptional and Epigenetic Dynamics Observed During Lineage Specification of Human Embryonic Stem Cells**

Abstract

Epigenetic regulation of gene expression is essential for faithful cellular specification during embryonic development. Directed differentiation of pluripotent human embryonic stem cells (hESCs), which maintain the ability to give rise to each cell type found within the human body, provides a tractable system to study both the epigenetic mechanisms that facilitate cellular transitions, and the transcription factors (TFs) that dictate these events. To understand molecular events associated with major lineage decisions, we performed comprehensive genomic profiling, including RNA-Sequencing, Chromatin Immunoprecipitation-Sequencing (ChIP-Seq) for six histone modifications and whole genome bisulfite-sequencing (WGBS) to interrogate DNA methylation levels, on three populations derived through directed differentiation of hESCs. Expression profiling detected signatures that resembled the three embryonic germ layers, namely ectoderm, mesoderm and endoderm. Integration of ChIP-Seq and WGBS data revealed widespread remodeling, predominantly at intergenic regions. To understand the impact of TF binding on epigenetic remodeling, we then complemented the epigenetic information with binding profiles for the pluripotency TFs OCT4, SOX2 and NANOG (O/S/N) in hESCs, and FOXA2 in the endoderm population. O/S/N binding was identified near pluripotency genes as expected, as well as regions that exhibited lineage specific remodeling during differentiation and are linked to later stages of development. We also discerned a novel epigenetic trend, in which H3K27me3 was unexpectedly gained at regions of low CpG density that exhibit high levels of DNA methylation in hESCs. These events overlapped with FOXA2 binding sites in the dEN that

lose DNA methylation. Notably, these events were detected near genes associated with later stages of development, such as *AFP*. We postulate that these FOXA2-associated epigenetic remodeling events lead to acquisition of a transient, facultative heterochromatic state necessary to foster efficient differentiation of subsequent stages. Integration of these data sets yielded an unprecedented perspective of the orchestrated transcriptional and epigenetic events that occur during cell state transitions. Future studies that compare epigenomic profiles of in vitro derived cell types to their primary counterparts may identify regulatory elements that are held in improper epigenetic states, and ultimately lead to improved differentiation protocols and the in vitro derivation of therapeutically relevant cell types.

## Table of Contents

## List of Figures

## List of Abbreviations

**5hmC** 5 hydroxymethylcytosine

**5mC** 5 methylcytosine

**bp** basepair

**CFP1** CxxC finger protein 1

**CGI** CPG island

**ChIP-BS-Seq** chromatin immunoprecipitation bisulfite-sequencing

**ChIP-Seq** chromatin immunoprecipitation-sequencing

**CpA** cytosine followed by adenine

**CpG** cytosine followed by guanine

**DamID** DNA adenine methyltransferase identification

**dEC** FACS-sorted HUES64-derived ectoderm

**dEN** FACS-sorted HUES64-derived endoderm

**dHep** HUES64-derived hepatoblast

**dME** FACS-sorted HUES64-derived mesoderm

**DHS** DNaseI hypersensitivity

**DMR** differentially methylated region

**DNMT1** DNA methyltransferase 1

**DNMT3A** DNA methyltransferase 3A

**DNMT3B** DNA methyltransferase 3B

**FPKM** fragments per kilobase of transcript per million mapped reads

**HAT** histone acetyltransferase

**H3K4me1** histone 3 lysine 4 monomethylation

**H3K4me2** histone 3 lysine 4 dimethylation

**H3K4me3** histone 3 lysine 4 trimethylation

**H3K9me3** histone 3 lysine 9 trimethylation

**H3K27ac** histone 3 lysine 27 acetylation

**H3K27me3** histone 3 lysine 27 trimethylation

**H3K36me3** histone 3 lysine 36 trimethylation

**H3K79me3** histone 3 lysine 79 trimethylation

**hESCs** human embryonic stem cell

**HMR** highly methylated region (61-100%)

**HMT** histone methyltransferase

**ICM** inner cell mass

**IMR** intermediately methylated region (11-60%)

**iPSC** induced pluripotent stem cell

**kb** kilobase

**KSR** knockout serum replacement

**lncRNA** long non-coding RNA

**MEF** murine embryonic fibroblast

**mESC** murine embryonic stem cell

**NDR** nucleosome depleted region

**NPC** neural progenitor cell

**OKSM** OCT4/KLF4/SOX2/cMYC

**OSN** OCT4/SOX2/NANOG

**PCC** Pearson Correlation Coefficient

**Pol II** polymerase II

**PRC2** polycomb recruitment complex 2

**PRE** polycomb recruitment element

**RNA-Seq** rna sequencing

**RPKM** reads per kilobase per million mapped reads

**SCNT** somatic cell nuclear transfer

**TET** ten eleven-translocation protein

**TF** transcription factor

**TSS** transcription start site

**UMR** unmethylated region (0-10%)

**WGBS** whole genome bisulfite sequencing

## Statement of Collaboration

Casey Gifford completed experiments included within this work independently, while the computational analysis of the epigenetic data was completed by Michael Ziller and the computational analysis of the RNA sequencing data was completed by Cole Trapnell. The majority of the interpretations are the result of collaboration between Casey and Michael. Casey wrote this thesis independently.

## Acknowledgements

I would like to thank members of the Rinn lab, especially Cole Trapnell and David Kelley.

Members of the Meissner lab, in particular Michael Ziller and Julie Donaghey.

My mentor Alex Meissner, for his patience and commitment to my scientific growth.

My family, for their love and support.

And finally my parents, for convincing me to do my homework.

Thank you.

# Chapter 1.

# Introduction

Excerpts from this chapter have been previously published. [1]

## 1.1 Introduction

The inner cell mass (ICM) of a blastocyst has the unique ability to give rise to any cell type found within the embryo proper [2]. This is possible due to the molecular plasticity of the pluripotent state, a feature that is gradually lost as a cell becomes specialized during embryonic development [3, 4]. The specification of embryonic lineages from the ICM is the result of a combination of intrinsic and exogenous stimuli, which leads to cells assuming the identity of one the three embryonic germ layers: ectoderm, mesoderm and endoderm [5].

Coordinated and timely propagation of epigenetic information, such as DNA methylation and the post-translational modification of histone proteins, cooperates with other mechanisms to guide cells along the path of lineage specification [6]. The term "epigenetic landscape," was coined to illustrate the concept that a cell's early choice of lineage dictates its future, as epigenetic mechanisms prevent lineage interconversion *in vivo*, leading to the gradual restriction of cell fate [2]. The derivation of human embryonic stem cells (hESCs) offered an exciting avenue to study these epigenetic decisions that dictate specification in a human context, as hESCs maintain the ability to differentiate towards the three embryonic germ layers *in vitro* [7]. This robust model system therefore creates a platform with which to study the epigenetic choices that promote cellular determination.

## 1.2 Epigenetic Regulation: DNA methylation

The term DNA methylation refers to cytosines that contain a methyl group on carbon 5, with the majority of DNA methylation occurring in a CG (CpG) dinucleotide context [8]. Throughout the human and mouse genome, 1 CpG is typically found every 100 base pairs (bp), and ≈80% of CpGs contain methylation in any given cell type [9-12]. Alternatively, CpG islands

(CGI) contain approximately 1 CpG every 10 bp and are typically void of DNA methylation [13]. They are found at 60% of human transcription start sites (TSS), and frequently overlap regions associated with embryonic transcription factors (TF) [13, 14]. The inclusion of a methylated cytosine within a regulatory element can prevent TFs from binding to DNA in some scenarios, which in turn limits transcription at that locus [1, 15]. Therefore, DNA methylation at certain regulatory elements is considered a stable form of gene repression, as it is prevalent in somatic cells at regulatory elements employed during embryonic development [8]. It is also involved in silencing regions associated with alternative lineages, that are presumably not required for maintenance of the specified somatic cellular identity exhibiting the DNA methylation [16]. In addition to its role in the regulation at specific genes, DNA methylation is also required for genomic stability, as disruption of DNA methylation patterns leads to aberrant chromosome segregation [17].

The addition of the methyl group is catalyzed by one of the three active DNA methyltransferases; DNMT1 [18], DNMT3a, or DNMT3b [19]. DNMT1 is generally responsible for copying the methylation pattern from the parent strand to the daughter strand during DNA replication [18], while DNMT3A/B exhibit de novo activity during cellular transitions [19, 20]. The deletion of DNMT3A/B results in embryonic lethality, suggesting that propagation of DNA methylation is essential for cellular specification [20]. Many reports have established that DNMT3A/B are responsible for DNA methylation at regions of genomic imprinting and satellite repeats [21-24], but the factor(s) responsible for catalyzing addition at specific gene regulatory elements during lineage specification and the timing of these events are not well defined [25].

CpG methylation must also be eliminated from regulatory elements, to alleviate repression and allow embryonic development to proceed. Recent studies indicated TF binding

and high DNA methylation tend to be mutually exclusive events [10, 11, 26], although the mechanism that induces the discrete loss during development remains somewhat elusive [27]. Recent identification of the TET enzymes, which catalyze the oxidation of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC) [28, 29], reveals a potential mechanism for DNA demethylation because DNMT1 does not recognize 5hmC [30].

### 1.3 Epigenetic Regulation: Histone Modifications

Nucleosomes, which consist of an octamer of histone proteins, provide another layer of epigenetic regulation of gene expression [31]. Post-translational modifications to the histone N-terminal tails, such as acetylation and methylation, have been reported for each histone [32]. These modifications regulate gene expression by affecting nucleosome organization within the nucleus, as well as polymerase II (Pol II) elongation during transcription [33, 34]. Histone lysine acetylation was one of the first post-translational modifications extensively studied [35], because early studies aimed at understanding the mechanism by which sodium butyrate induces cellular transitions found it led to acetylation of H3 and H4 [36, 37]. This modification was shown to prevent chromatin compaction [38], and its discrete localization was associated with gene expression [39].

With evidence that histone acetylation was essential for gene regulation, studies then focused on substrate specificity of histone acetyltransferases (HATs) to understand if acetylation was randomly distributed throughout the histone protein. Studies in yeast and drosophila confirmed that it was non-random, and that HATs target specific residues [40, 41]. For example, H3K27ac is a commonly studied histone modification, which demarcates a subset of active promoters and enhancers [42, 43].

Numerous subsequent studies devised to understand gene regulation via histone modifications suggested that multiple histones/residues/modifications need to be considered in tandem to infer regulatory impact of the surrounding chromatin landscape [44]. Additional consideration must be given to modifications such as lysine methylation, which correlate both with, and against transcriptional activity according to ChIP-Sequencing (ChIP-Seq) studies that established genome-wide maps of histone modifications [45]. A repressive role was reported for H3K9me3 and H3K27me3, which were enriched at various loci known to exist in a closed or heterochromatic state, such as centromeres, repeat elements and the inactivated X chromosome [46]. H3K9 methylation is most commonly associated with repeat elements, but promoter association was revealed by studying the correlation between DNA hypermethylation and cancer [47, 48]. Gene silencing mediated by H3K27me3 [46] is alternatively associated with regions of high CpG density [49]. Addition of H3K27me3 by the Polycomb repressive complex (PRC2) mediates repression by promoting a compact chromatin structure [50, 51].

A divergent role was reported for lysine methylation, as H3K4me was first associated with active transcription in *Saccaromyces cerevisiae* [52], and then extended to vertebrates [53]. Association with distinct genomic features was also reported as H3K4me3 was found at promoters while H3K4me1 was enriched at distal enhancers [23, 52, 54]. At promoters, H3K4me3 allows recruitment of nucleosome remodeling enzymes and HATs, which together create chromatin architecture amenable to transcription initiation [55-57]. H3K4 methylation is catalyzed by SET-domain containing histone methyltransferases (HMT), which maintain an opposing function to PRC2 [58]. Disruption of various SET domain containing-complex subunits causes severe embryonic defects and prevents efficient differentiation of mESCs *in vitro* [59-61]. Marks such as H3K36me3 and H3K79me3 are also associated with active transcription, but they

are enriched over actively transcribed gene bodies [49, 62]. In conclusion, histone methylation is utilized in various contexts.

### 1.4 Integration of Epigenetic Mechanisms

Our ability to assess the regulatory function of epigenetic mechanisms was vastly improved with the creation of sequencing-based approaches. Rather than being limited to the interrogation of a few specific loci, sequencing provided a genome-wide enrichment profile for the protein or modification of interest. As more genome-wide studies were conducted, the relationships between coexisting epigenetic mechanisms and expression were defined, and are now commonly referred to as epigenetic states [42]. For example, it is now understood that some histone modifications maintain a mutually exclusive relationship with DNA methylation, such as H3K4me3 and H3K27me3 [14, 63]. This is supported by *in vitro* experiments that showed Dnmt3a activity was stimulated by an unmethylated H3 tail [64]. Though a positive correlation is exhibited between DNA methylation and Suv39h, a HMT that catalyzes H3K9me3 [48]. DNA methylation maintenance at pericentric satellite repeats required this HMT, and it reportedly interacts with all three of the DNMTs [23, 48, 65]. Gain of DNA methylation at the OCT4 promoter during differentiation also requires G9a, another H3K9me3 HMTase, to promote DNMT3A/B-directed stable silencing [66, 67]. Expression of OCT4 decreases with the addition of H3K9me3 in this study, which suggests DNA methylation serves as a secondary form of silencing. DNA methylation is also found in concert with H3K36me3 over actively transcribed gene bodies, but DNA methylation appears to be propagated independent of H3K36me3 enrichment given that depletion of SETD2, a H3K36me3 HMT, did not alter DNA methylation levels over gene bodies [68].

The cooperation between multiple histone modifications became a major topic of interest with the observation that H3K27me3 and H3K4me3 were commonly found at regions of high CpG density in ESCs that were near regulators of embryonic development, and termed a bivalent domain [69, 70]. It was hypothesized that this combination facilitates swift activation during specification [71]. Similar to bivalent promoters, the combination of H3K4me1 was enriched with H3K27me3 at putative distal regulatory elements associated with transcriptional silencing, and termed poised domains [43, 72]. These elements were reported to then transition during development to include H3K27ac enrichment rather than H3K27me3, in addition to H3K4me1 localization when the regulatory element was activated [43].

While these studies suggested genomic overlap of histone modifications, some technical limitations originally prevented more discrete conclusions regarding the extent of the co-occurrence. It was difficult to distinguish if these modifications occurred on different alleles leading to their putative overlap in the maps. Similarly, it was unclear if they occurred on the same nucleosome, which includes two molecules of each histone protein. It also remained possible that heterogeneity existed within the population, and the co-enrichment did not occur in one cell. More recent studies revealed that H3K4me3 and H3K27me3 do not coexist on the same histone tail but can occur on different H3 tails in one nucleosome [73]. This correlated with earlier work suggesting PRC2 activity is inhibited by the presence of H3K4 methylation [74].

Our understanding of the interplay between DNA methylation and DNA-protein interactions also improved dramatically with the advent of whole genome bisulfite sequencing (WGBS), which allowed the interrogation of most CpG dinucleotides encoded within a genome [10, 11]. Integrating this data with histone modification and protein binding maps revealed regions of low DNA methylation at intergenic sites that frequently overlapped with DNaseI

hypersensitivity (DHS), a symbol of open or accessible chromatin, and/or displayed an active enhancer signature defined by H3K4me1 and p300 enrichment [11]. A localized depletion of DNA methylation was also detected at regions of TF binding [10, 11], though the factors that catalyze this depletion at these sites was not fully elucidated.

WGBS sequencing also confirmed a notable characteristic of the stem cell methylome; non-CpG methylation. DNA methylation presumed to be of regulatory significance is commonly studied in the CpG context, given historical studies that suggested greater than 90% of methylcytosine was detected in this context [9]. But profiling of mESCs initially revealed that CpA methylation was also prevalent, and likely catalyzed by DNMT3A [75]. Many years later, WGBS confirmed that CpA methylation was detectable in mESCS and hESCs, and also found that depletion of DNMT3A reduced the levels of CpA methylation in hESCs [10, 11, 76, 77]. More recent reports of WGBS from many stages of neural development also reported non-CpG methylation in the frontal cortex, which correlated with reactivation of *DNMT3A* expression [77, 78].

## 1.5 Directing Epigenetic Remodeling

While sequencing-based studies established that histone modifications exhibit discrete localization, the mechanisms involved in targeting the addition and removal of histone methylation at specific genomic loci during cellular transitions remain somewhat unclear. It has been suggested that the CxxC finger protein CFP1 directs SET1 to promoters leading to enrichment of H3K4 methylation [79], which supports gene activation. An additional report suggested the TF USF1 recruits SET1A to genomic regions, which stimulated activation of target

genes during differentiation [80]. But the factors responsible for widespread recruitment of HMTs to genomic locations during specification are mostly unknown.

The mechanism of recruitment of complexes that catalyze the addition of repressive histone modifications is similarly unresolved in higher eukaryotes. While the *Drosophila melanogaster* genome encodes polycomb recruitment elements (PREs), which are responsible for recruiting the PRC2 complex to specific genomic sites to maintain silencing [81], similar sequences have been more difficult to find in mammals. As previously mentioned, H3K27me3 is associated with regions of high CpG density in vertebrates, but the additional underlying sequence context that recruits PRC2 to these regions genome-wide is an area under heavy investigation. In an attempt to understand PRC2 recruitment, an activating TF motif was deleted from an ectopic reporter sequence with a high GC content. This deletion lead to loss of the activating modification H3K4me3 and gain of H3K27me3 in mESCs [82], suggesting that in the absence of an activating TF that recruits HMTs to catalyze addition of a histone modification such as H3K4 methylation, H3K27me3 is gained by default. A subsequent study examined H3K27me3 dynamics during murine neural progenitor cell (NPC) differentiation, and found that inclusion of CoREST and SNAIL binding sites in an ectopic reporter also lead to high H3K27me3 enrichment [83]. Many studies, like those highlighted here, that attempted to establish recruitment mechanisms focused on a small number of loci or cell types, therefore this area of epigenetic regulation of gene expression requires further investigation.

### 1.6 Signaling Pathways Associated with Chromatin Remodeling

Numerous signaling cascades converge on the nucleus during development to direct epigenetic remodeling that influences transcriptional regulation. For example, proper TGFβ

signaling gradients are essential for cellular viability and differentiation [84]. Activation of TGFβ receptors located at the cell surface catalyzes a phosphorylation cascade that leads to SMAD protein interactions in the cytoplasm and their subsequent relocation to the nucleus [85]. Interestingly, high cell density of various cell lines prevented SMAD translocation to the nucleus *in vitro*, suggesting extrinsic factors in addition to morphogen gradients that may contribute to TGFβ signaling [86].

The SMAD1/5/8 complex is activated via BMP GDF ligands, while Activin/NODAL-associated SMAD2/3 complex [87, 88]. Once transported inside the nucleus, the SMAD complex interacts with additional transcription factors to bind the SMAD DNA binding element and activate target gene expression [89]. Work in *X. laevis* showed that the SMAD2/3 complex interacts with Mix and FoxO family transcription factors to activate genes such as *GSC* and *p21* [90, 91]. This complex activates gene expression by recruiting the HAT p300 and Brg1 to target loci [92]. The SMAD2/3 complex also represses gene expression when it associates with ATF3 [93], suggesting it has various roles in gene regulation. Recent ChIP-Seq studies also showed that SMAD1 binding overlapped with GATA1/2 in K562 cells, and could be re-directed to new binding sites with ectopic expression of CEBPα [94]. Ectopic expression of *MyoD* in mESCs also provided evidence that SMAD3 can be re-directed to new binding sites by this TF [95]. Collectively, these studies suggest that master TF regulators of cell fate can direct SMAD binding which putatively induces epigenetic remodeling and gene activation.

WNT/β catenin signaling has also been show to direct chromatin remodeling. WNT ligands bind to Frizzeled and LRP receptors on the cell surface, allows β catenin's translocation to the nucleus where it interacts with TCF/LEF transcription factors to promote gene activation [96]. In addition to an overlap between GATA1/2 and SMAD1 previously mentioned,

Trompouki and colleagues also found that TCF7L2 binding overlapped with SMAD1/GATA1/2 binding, suggesting an overlap in WNT and BMP signaling pathways [94]. In some scenarios, the C terminus of β catenin directly interacts with the HMTs MLL1/MLL2 to stimulate H3K4 methylation, which leads to expression of genes such as *c-Myc* [97]. Modulation of these pathways through extracellular binding of ligands therefore represents a mechanism that links developmental signaling cues with nuclear remodeling.

## 1.7 Transcription Factors and the Epigenome

Unlike the pluripotent epigenome, somatic cells retain a restricted epigenetic landscape that does not readily respond to signaling cues in the same manner as embryonic stem cells. Nuclear remodeling can alternatively be induced by ectopic expression of TFs. With the report that ectopic expression of MyoD could induce a fibroblast to acquire myotube-like characteristics, many began to explore the ability of lineage specific TFs to engage with and remodel the epigenome in an ectopic context [98]. TF-mediated reprogramming experiments were complimented by studies using cell fusion and somatic cell nuclear transfer (SCNT), which demonstrated the ability to revert somatic nuclei to pluri- and totipotency, respectively [99]. However, arguably the most notable advancement in the understanding of TF capabilities came with the report of successful reprogramming of a somatic cell to an induced pluripotent state by ectopic expression of OCT4, SOX2, KLF4 and c-MYC (OSKM). This suggested that at some frequency, TFs could remodel the epigenetic state of a restricted cell type back to an open, pluripotent state.

While the reprogramming process does result in the derivation of a seemingly pluripotent population, it exhibits limited efficiency suggesting that only a fraction of the somatic genome is

amenable to epigenetic remodeling as a result of ectopic OSKM expression. To define the epigenetic landscape that determines susceptibility to OCT4 binding and directed remodeling in somatic cells, Taberlay and colleagues found that putative enhancer regions enriched for H3K27me3 permit OCT4 binding if a nucleosome-depleted region (NDR) is nearby [100]. The NDR is significant in this context, as it's creation is a necessary step in the regulation of transcription and suggests that additional proteins have already bound to the site and created an accessible chromatin landscape [101].

Ectopic expression of *OCT4* in human fibroblasts confirmed that this TF could bind to the NDR within the H3K27me3-enriched enhancer of MYOD, subsequently leading to the remodeling of the promoter to a bivalent state. A complimentary report suggested that OCT4 could promote establishment of NDRs at TSSs, exclusively in the absence of DNA methylation [102]. These results suggest that enrichment of H3K27me3 within distal regulatory elements provides a degree of local permissiveness to TF binding and chromatin remodeling. c-MYC, a non-essential reprogramming factor, is categorized as a transcriptional pause release factor, and also likely exhibits a limited ability to engage heterochromatin. It associates with the Pol II pre-initiation complex, which cannot assemble without pre-existing H3K4 methylation [57, 103, 104].

In contrast to the restricted capacity of OCT4 and c-Myc, TFs such as FOXA1 exhibit a relatively unique ability to access and remodel target chromatin in an ATP-independent fashion *in vitro* [105-107]. A group of factors with this ability "pioneering transcription factors," defined by their ability to remodel heterochromatic regions subsequently leading to open chromatin and transcriptional activation [108]. Consequently, many TF combinations that include pioneer factors have been successfully used to induce a somatic cell to assume a new, alternative identity

[98]. Pioneering factors such as FOXA1 and GATA4 perform their chromatin remodeling functions as monomers [109-111], unlike OCT4, which requires dimerization with SOX2 for functionality in a pluripotent context [112, 113]. Further insight regarding the role of dimerization during reprogramming was provided by a recent demonstration that SOX17 could be transformed into a reprogramming factor by changing specific residues that promoted its interaction with OCT4 [114]. While the DNA binding profile of the amended dimer was not directly interrogated, the ability of multiple Sox family members to serve as iPSC reprogramming factors suggests that OCT4 may dictate DNA binding specificity while SOX2 recruits HATs, such as p300 [115, 116]. The binding site preference for OCT4 is related to available co-factors, given that OCT4's interaction with SOX17 during differentiation leads to regulation of different genes compared to the OCT4/SOX2 dimer that is active in ESCs [117].

Interestingly, Hiryai et al recently affixed the transactivation domain (TAD) of MyoD to OCT4, and examined this engineered factor's ability to promote the reprogramming of MEFs [118]. Indeed they found that substituting the endogenous form of OCT4 with the TAD fusion construct accelerated the appearance of OCT4 positive colonies during reprogramming. This suggests that the endogenous form of OCT4, in combination with the other reprogramming factors (SKM), has limited inherent ability to access and remodel heterochromatic regions independently, and that this constraint can be relaxed by introduction of the TAD. Potentially contributing to this limitation is the stability with which these factors bind to DNA given evidence that forkhead domain-containing factors remain bound to mitotic chromosomes [119], while Pou family members do not [120, 121]. However this function, known as bookmarking, is associated with pioneering transcription factors [122]. The ability to interact with DNA and cause epigenetic remodeling is therefore variable between different classes of TFs.

## 1.8 Specific Aims

At the inception of my thesis work, genome-wide profiles were available for various histone modifications, as well as transcription factors, derived from both pluripotent and somatic cell types. Basic associations, such as H3K4me3 enrichment at promoters and H3K4me1 at putative enhancers, were well established. Many cell type-specific regulatory elements and general genome-wide remodeling trends had also been reported. However, these observations were made by comparing profiles of unrelated cell types, such as pluripotent stem cells and immortalized cell lines, limiting our understanding of how a genomic region transitions to a new epigenetic landscape during development and differentiation.

Consequently, our goal was to create an *in vitro* system that would reveal transient epigenetic states associated with lineage decisions, similar to those that dictate embryonic development. We additionally hypothesized that the creation of this system would offer the opportunity to identify individual factors that induce these transitions, and assess whether these factors acted in a lineage specific manner. Therefore, we chose to use directed differentiation of hESCs to illuminate the epigenetic events that facilitate silencing of the pluripotent network and simultaneous activation of lineage specific programs.

By modulating the signaling environment that maintains pluripotency of human embryonic stem cells, we induced hESC differentiation to three distinct populations that resembled ectoderm, mesoderm and endoderm using 2-dimensional (2-D) directed differentiation (Chapter 2). We then performed RNA-Sequencing (RNA-Seq), ChIP-Seq for six histone modifications and WGBS on each population, as well as the starting hESCs (Chapter 2), to define transcriptional and epigenetic dynamics shared by all lineages, as well as those events that are lineage specific. Given that the gain of DNA methylation is associated with stable repression

of gene regulatory elements, we first focused our analysis on DNA methylation dynamics to identify regulatory elements that are silenced during specification, and to understand the timing and consistency of these events across each germ layer (Chapter 3). This analysis was complimented by OC4, SOX2 and NANOG (O/S/N) binding profiles in hESCs, which we created to first define elements associated with the core pluripotency network and subsequently to understand the timing and mode involved in their transition to a stably repressed state. To understand the process by which lineage specific gene networks are activated, we then integrated the DNA methylation and ChIP-Seq data to identify regions that transition to a euchromatic state during differentiation (Chapter 4). We additionally created a binding profile for FOXA2 in the endoderm population to expand our understanding of its role during specification (Chapter 4). Collectively, this extensive data set uncovers epigenetic dynamics that accompany lineage specification *in vitro*.

# Chapter 2.

# Transcriptional and Epigenetic Profiling of hESC-derived Populations

Work presented in this chapter is previously published. [123]

## 2.1 Rationale

Cellular specification involves a complex series of molecular decisions that are based on interpretation of a myriad of signaling cues. While previous genetic studies have established that epigenetic remodeling is essential for specification, the precise genomic locations of epigenetic remodeling events that facilitate human embryonic lineage specification have not been widely identified. The first goal of my thesis was to derive populations that represent each embryonic germ layer through directed differentiation of hESCs, and then subject these populations to epigenomic profiling. We believed that this would allow us to identify regions of regulatory significance that are associated with specification, and reveal genome-wide trends associated with lineage decisions.

## 2.2 Derivation of Three Populations that Resemble Embryonic Germ Layers

We chose to work with the male line, HUES64, which was expanded and maintained on murine embryonic feeders (MEFs) in the presence of knockout serum replacer (KSR) and basic fibroblast growth factor (bFGF, 10ng/mL). This line actively stained positive for pluripotent markers such as NANOG and OCT4 when cultured in feeder-free conditions (**Figure 2.1, left**), which we defined as matrigel coated plates and mTeSR medium. After allowing the cells to acclimate to this culture environment, we used one of three approaches to induce direct differentiation of distinct lineages. HUES64 readily differentiated into a neuroectoderm-like progenitor population positive for SOX2 and PAX6 by adding medium containing KSR and antagonists of TGFβ, WNT and BMP signaling (**Figure 2.1, top right**) [124]. We altered the previously published protocol due to reports that sustained WNT signaling can prevent neural differentiation [125]. Alternatively, canonical mesoderm markers, such as GATA2 (**Figure 2.1,**

**middle right**) and HAND2, were induced by switching the culture environment to low serum (0.5%) and adding ACTIVIN A, BMP4, VEGF and FGF2 [126]. Differentiation towards a definitive endoderm-like fate, positive for markers such as SOX17 and FOXA2 (**Figure 2.1, bottom right**), was induced by again switching to a lower serum environment (0.5%) and adding ACTIVIN A and WNT3A [127, 128].

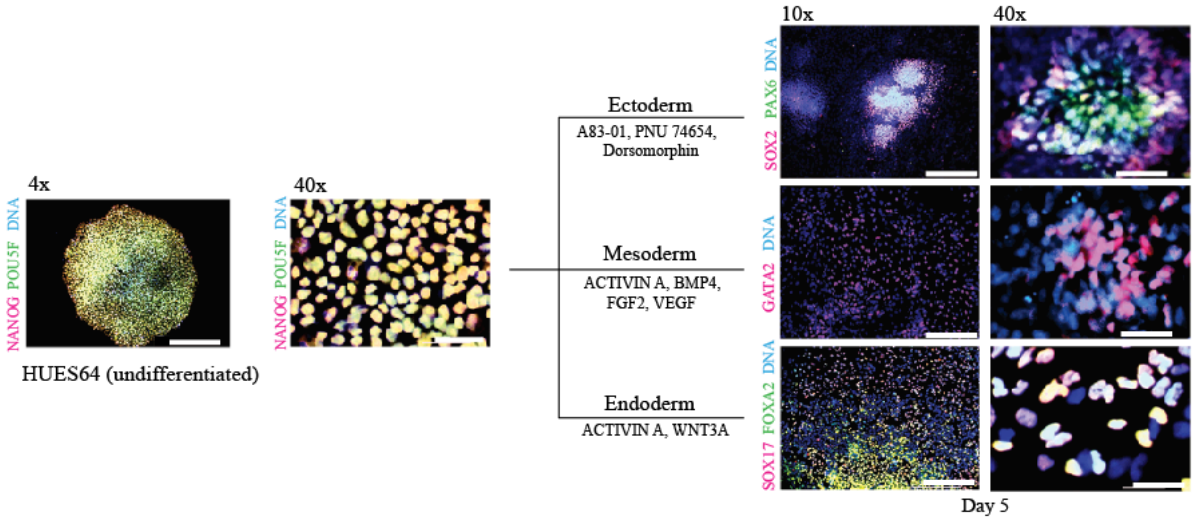



**Figure 2.1** Left: Low (4×) and high (40×) magnification overlaid immunofluorescent images of the undifferentiated hESC line HUES64 stained with OCT4 and NANOG antibodies. Right: Directed (two-dimensional) differentiation conditions were used to generate representative populations of the three embryonic germ layers. Cells were fixed and stained after 5 days of differentiation with the indicated antibodies. Representative overlaid images at low (10×) and high (40×) magnification are shown. DNA was stained with Hoechst 33342 in all images. Scale bars, 200 μm (4×), 100 μm (10×), and 30 μm (40×).

To understand the transcriptional dynamics within the first five days of differentiation, we measured the expression of 541 selected genes, including many developmental transcription factors and lineage markers, using NanoString profiling, at 24-hour intervals during differentiation towards each respective germ layer. NanoString is a semi-quantitative mRNA profiling method that measures mRNA molecules independent of PCR amplification [129].

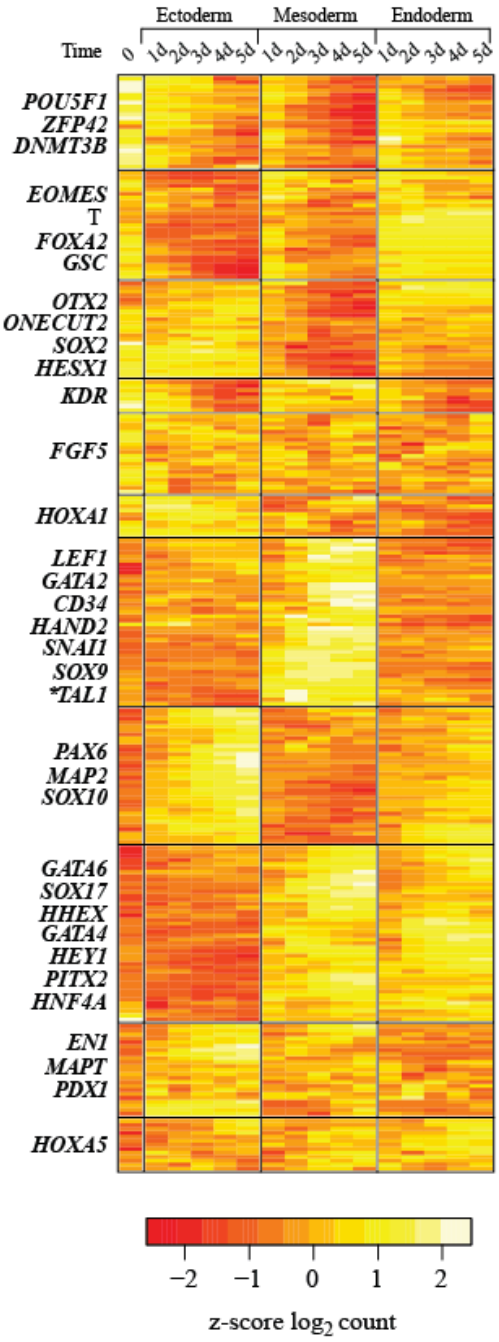



**Figure 2.2** NanoString nCounter expression data (Z score $\log_2$ expression value of two biological replicates) for a time course of *in vitro* differentiation using the conditions shown in (2.1). 541 genes were profiled, and 268 genes that changed by more than 0.5 are displayed. Selected genes are shown on the left for each category that was identified based on hierarchical clustering.

We found that 268 of these genes exhibited expression changes compared to hESCs during the first five days of differentiation (**Figure 2.2)**. Mesendodermal genes, such as *EOMES, T, FOXA2* and *GSC*, were upregulated at 24 hours of mesoderm and endoderm induction, but not ectoderm differentiation (**Figure 2.2**). GSC expression decreased within 48 hours of differentiation in the mesoderm-like population, while the expression level was maintained in the endoderm population (**Figure 2.3A)**. *EOMES* and *FOXA2* expression was also maintained in the endoderm population accompanied by upregulation of *GATA6, SOX17* and *HHEX* (**Figure 2.2**). After transient upregulation of mesendodermal markers, activation of mesodermal markers such as *GATA2, HAND2, SOX9* and *TAL1* was detected specifically in the mesoderm conditions (**Figure 2.2, 2.3B)**. By day 5, expression of genes such as *PAX6*, *SOX10* and *EN1* were detected in the ectoderm popuation (**Figure 2.3C**). Multidimensional scaling confirmed that at 24 hours, the mesoderm population is very similar to the endoderm, while the ectoderm population has already moved in an alternative direction (**Appendix, Figure S1**).

To confirm that these cells were in fact differentiated, we also examined expression of pluripotency related genes. We found that *OCT4* and *NANOG* expression was maintained in our endoderm population (**Figure 2.3D, E**). This is consistent with prior studies indicating that OCT4 and NANOG expression is detected during the course of early endoderm differentiation and supports NANOG's suggested role in endoderm specification [130]. *SOX2* expression was downregulated in mesoderm and to a lesser degree in endoderm, but maintained at high levels in the ectoderm population (**Figure 2.3F**), while *ZFP42* (REX1) was similarly down regulated in all three lineages. Based on these results, we selected day five as the optimal time point to capture early regulatory events in differentiated populations representing all three germ layers.

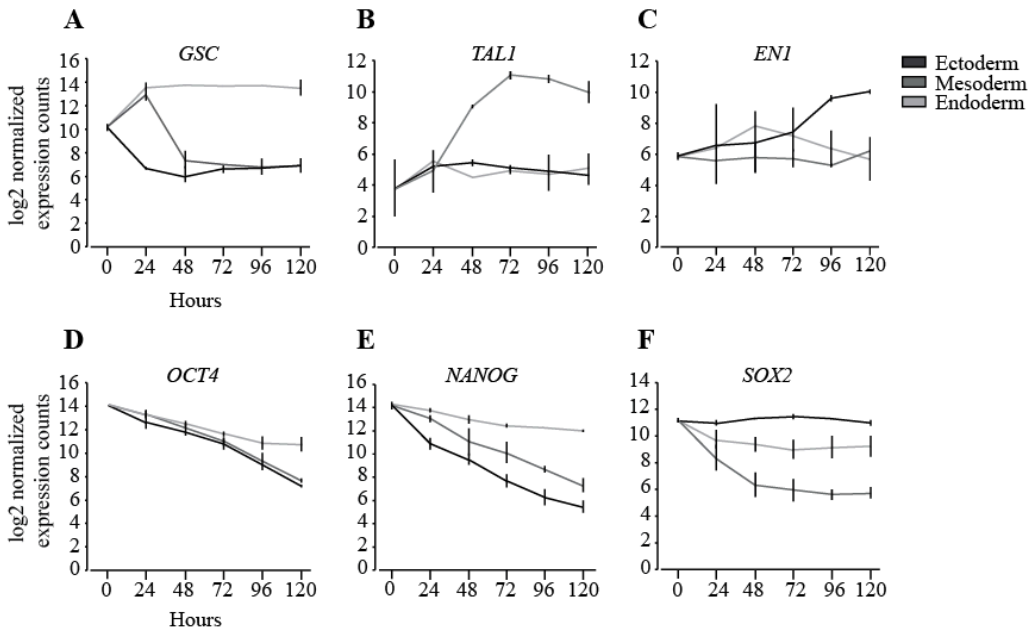



**Figure 2.3 A-F** The average $\log_2$ expression values from two biological replicates of six genes is shown. Error bars represent 1 standard deviation (SD). If no error is evident, SD < 0.2 $\log_2$ expression units.

We confirmed that these populations indeed represented a precursor stage for each respective lineage by inducing them to differentiate further, which resulted in upregulation of genes such as *OLIG2* and *SST* in the ectoderm [131], *TRPV6* in the mesoderm [126], and *AFP* and *HGF* in the later endoderm populations [132] (**Appendix, S2**).

To reduce heterogeneity, we used FACS to enrich for the desired populations based on previously reported surface markers (**Appendix, S3**). Populations isolated by FACS are herein referred to as dEC for the ectoderm, dME for the mesoderm and dEN for the endoderm. Expression analysis of the sorted populations confirmed enrichment for the desired populations (**Figure 2.4**). For example, profiling of the unsorted ectoderm population suggested that *SOX17* was expressed during ectoderm differentiation. Sorting of the day 5-ectoderm population revealed that while the CD56h/CD326l population maintained expression of *PAX6, EN1* and *SOX2*, expression of *SOX17* was reduced to background levels (**Appendix, S4**).

Immunofluorescent staining confirmed that the SOX17 and PAX6 positive populations were mutually exclusive (**Appendix, S4**). We concluded that sorting was necessary to facilitate our downstream analysis.

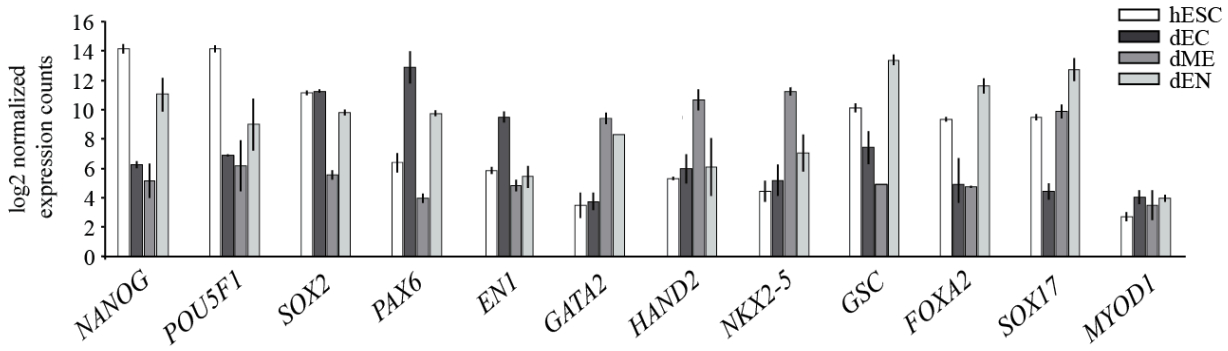



**Figure 2.4** NanoString nCounter profiling of FACS-isolated ectoderm (dEC), mesoderm (dME), and endoderm (dEN). Expression levels for MYOD1 (right) are included as a negative control. The average $\log_2$ expression value of two biological replicates is shown. Error bars represent 1 SD. If no error is evident, SD < 0.2 $\log_2$ expression units.

2.3 Global Expression Analysis

We next expanded on our NanoString expression profiles by performing strand specific RNA-Seq on poly-A fractions from each of the FACS-isolated populations and undifferentiated HUES64 (**see Methods**). Hierarchical clustering using the Jensen-Shannon metric based on the global expression profiles of each cell type revealed that the dEN and dEC were more similar to each other than to dME or hESCs (**Figure 2.5A**). This was unexpected given that the dME and dEN populations are putatively derived through a common mesendoderm precursor stage based on NanoString profiling while the dEC did not upregulate similar markers (**Figures 2.2**). Overall, 14,196 RefSeq-defined coding and non-coding transcripts were expressed (**Figure 2.5B**, Fragments per kilobase of transcript per million mapped reads (FPKM: >1) in at least one of the populations (≈38% of transcripts measured),

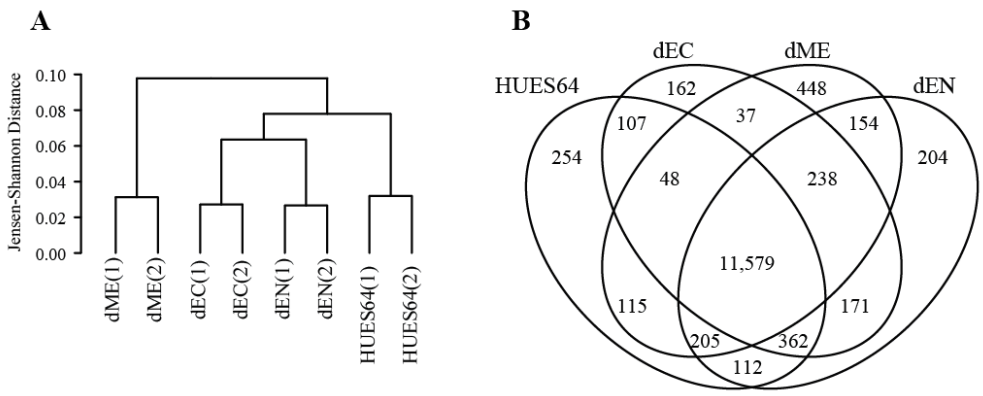



**Figure 2.5 (**A**)** Hierarchical clustering of global gene expression profiles as measured by strand-specific RNA-seq for biological replicates shown as a dendrogram. Pairwise distances between the replicates were measured using the Jensen-Shannon distance metric. (B) Venn diagram illustrating unique and overlapping genes with expression FPKM > 1.

with 11,579 (81.6% of the total number of transcripts expressed within our cell types) being expressed in all four populations (**Figure 2.5B**).

Examination of the overlap of genes expressed in each population revealed that the dME population exhibited expression of the largest number of unique genes (n=448, **Figure 2.5B**), such as *RUNX1* (FPKM: 3.4) and *HAND2* (FPKM: 17.8). Genes unique to pairs of the differentiated cell types also revealed that dEC and dME had the least in common (n=37, **Figure 2.5B**), while the dEC and dEN had the most number of transcripts in common (n=171**, Figure 2.5B**), consistent with our clustering analysis. Genes such as *PAX6* (dEC FPKM: 25.9, dEN FPKM: 5.6) and *NKX6.1* (dEC FPKM: 2.3, dEN FPKM: 3.3), which are each required for both brain [133] and pancreas development [134], were expressed in both the dEC and dEN. Canonical markers of embryonic development, such as *FOXA2* (FPKM: 12.7) in the dEN and *EN1* (FPKM: 5.8) in the dEC were restricted to their expected germ layers in the populations that we profiled.

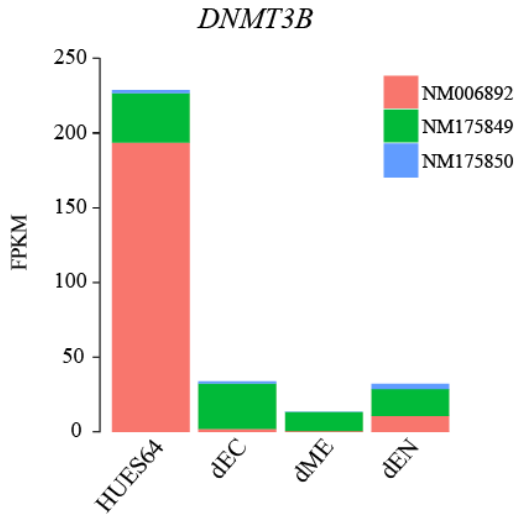



**Figure 2.6** Differential splicing and downregulation of *DNMT3B* during directed differentiation to each cell type (FPKM).

Notably, we also identified 1,296 splicing and alternative promoter usage events within our populations (FDR=5%) [135]. For example, we detected expression of multiple isoforms of *DNMT3B* ($p=5x10^{-5}$). Expression of *DNMT3B* isoform 1 (NM_006892) was restricted to the undifferentiated hESCs (FPKM: 214.3), while the differentiated cell types predominantly expressed an alternative isoform, *DNMT3B* isoform 3 (NM_175849), though at much lower levels compared to hESCs (**Figure 2.6**). *DNMT3B3* is expressed in various cancers and adult cell types, but our results suggest that this switch coincides with the exit from the pluripotent state, regardless of the specified lineage [136-138]. This splicing event also occurs during murine embryonic development [139] and it is particularly intriguing because this isoform does not maintain DNA methyltransferase activity [140].

Extending our analysis to include non-coding elements, we identified 3,219 long non-coding RNAs (lncRNA) that were longer than 200bp, spanned an intron and were expressed at >0.25 FPKM within our four populations. lncRNAs appear to be expressed at lower levels than

protein coding genes, hence the lower expression threshold [141]. Unlike protein-coding genes, few of lncRNAs identified in our cell types include CpG islands in their promoter, which was confirmed by an additional report on hESC differentiation [142]. Though one divergent example is the non-coding element *megamind,* which includes a CpG island at its promoter. This transcript was originally identified in an analysis designed to uncover novel lncRNAs involved in zebrafish development [143]. The zebrafish transcript was expressed in the retina and brain at 28 hours post-fertilization, and morpholino-facilitated depletion during development resulted in a smaller brain, and loss of NeuroD-positive neurons in the retina, among other phenotypes. Interestingly, this lncRNA is expressed in our dEC, suggesting conservation of this element in the human ectoderm lineage. The expression dynamics presented here suggest that we have isolated three distinct populations that represent different embryonic germ layers.

## 2.4 Integrative Analysis of Epigenetic Dynamics

To gain a more complete picture of the underlying molecular dynamics that dictate specification of the three germ layers, we collected approximately 10 million cells of the respective dEC, dME and dEN populations, as well as HUES64, and subjected to them ChIP-Seq (data is publicly available through the NIH Roadmap Epigenomics Project data repositories: http://www.roadmapepigenomics.org/). We chose the modifications H3K4me1 and H3K4me3 to demarcate open enhancers as well as promoters, H3K36me3 for transcribed gene bodies, H3K27me3 and H3K9me3 to identify repressed regions and H3K27ac to identify active enhancers. Three genomic loci and their associated epigenetic landscape are depicted in **Figure 2.7.**

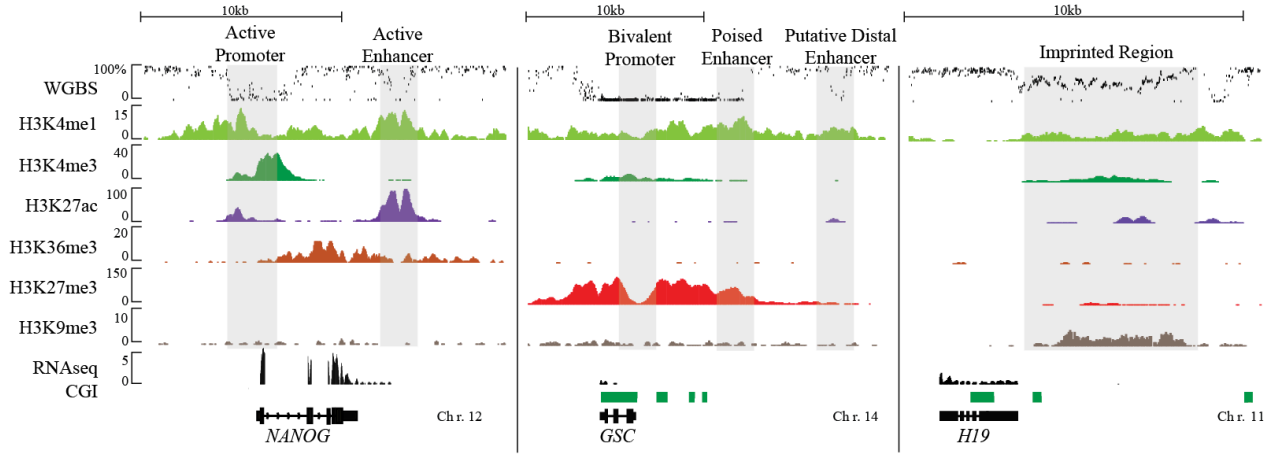



**Figure 2.7** WGBS (% methylation), ChIP-Seq, and RNA-Seq for the undifferentiated hESC line HUES64 at three loci: NANOG (chr12: 7,935,038-7,957,818), GSC (chr14: 95,230,449-95,250,241), and H19 (chr11: 2,015,282-2,027,359). CGIs are indicated in green.

After completing our basic quality control (see methods), we focused our analysis on previously identified informative chromatin states associated with various types of regulatory elements [42, 43], including the following specific combinations: H3K4me3+H3K27me3, H3K4me3+H3K27ac, H3K4me3, H3K27me3+H3K4me1, H3K4me1, H3K27ac+H3K4me1, H3K27me3 and H3K9me3. To identify genomic regions of enrichment for each histone modification, we segmented the genome in to non-overlapping windows and calculated enrichment based on the number of unique reads whose midpoint was located within the window of interest, compared to the enrichment within the corresponding window in the input control sample. Regions that exhibited >3 fold enrichment over background at a significance level of $p < 10^{-5}$ were considered enriched [144]. Enriched windows occurring within 850bp of each other were merged together in to one region, whereas regions smaller than 400bp were excluded. Then, regions enriched for multiple histone modifications were identified. In addition, we segmented the genome-wide DNA methylation levels in to three CpG methylation states [11]: highly

methylated regions (HMRs: >60%), intermediately methylated regions (IMRs: 11-60%) and unmethylated regions (UMRs: 0-10%). The latter differs from the generally high methylation level throughout the majority of the genome and likely indicates functional importance as previously suggested [11].

We next assigned each genomic region to one of the resulting states (**Figure 2.8A**). Of the identified epigenetic states, dynamic regions that were H3K9me3-enriched or HMRs covered the most base pairs (**Figure 2.8B)** and the combination of H3K4me3 and H3K27me3 exhibited the highest CpG content (**Figure 2.8C**), as expected given the close association between H3K27me3 and high CpG density [49]. Regions of open chromatin, enriched for H3K4me3 exhibited the highest median expression value (**Figure 2.8D**). The size of dynamic regions spanned 400 bp to 10 kilo base (kb) (**Figure 2.8E**) and many dynamic regions were distal, as 48.8% were located greater than >50 kb from the nearest TSS (**Figure 2.8F**).

Across the four cell populations, we identified 268,062 genomic regions that cover a total of 450,058,678 bp, that were enriched for at least one of the 8 chromatin states and/or classified as an IMR or UMR in at least one of the cell types. During differentiation, 157,443 regions changed their state in at least one of the populations (**Figure 2.9**). Interestingly, overall we found that ≈62% of all epigenetic state changes are not directly linked to transcriptional changes based on the expression of the nearest gene using RNA-Seq (change >1.5 fold).

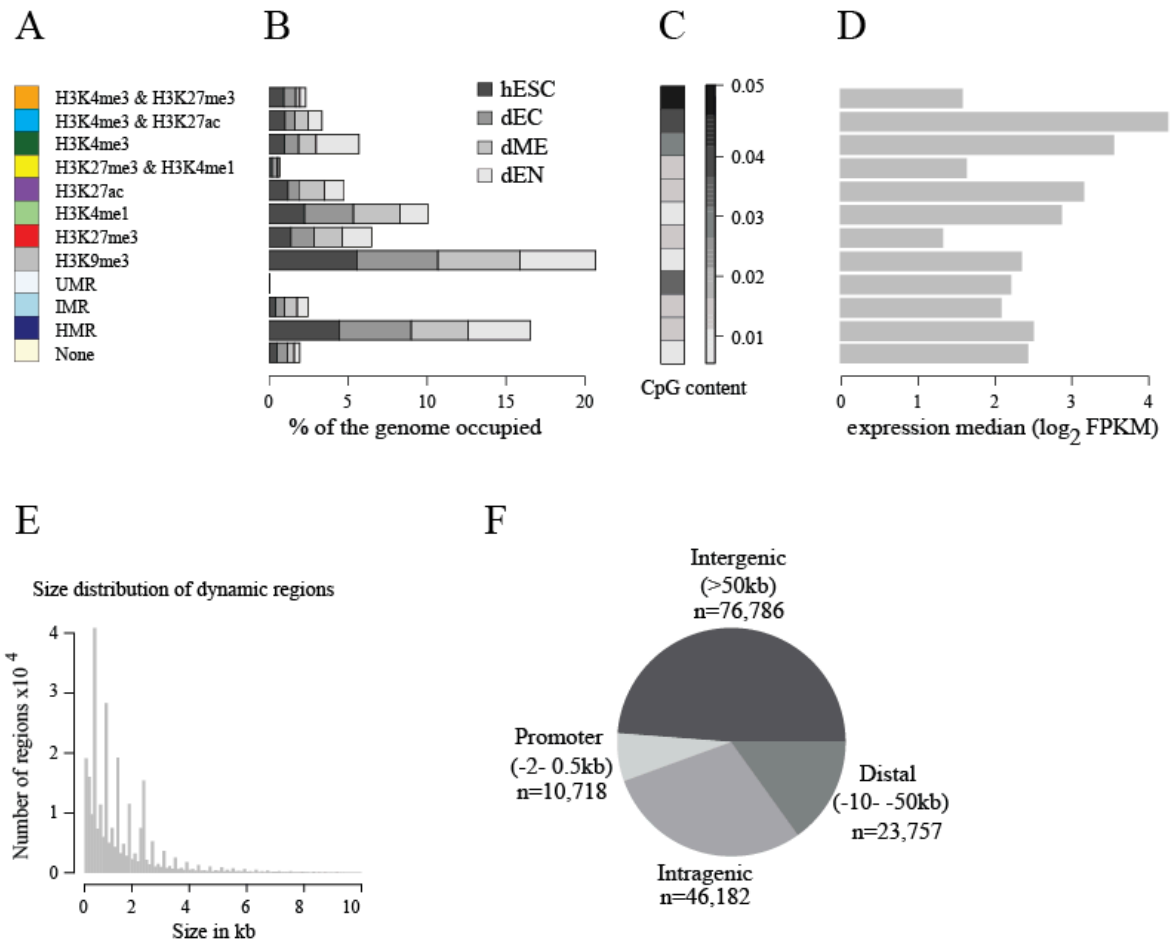



**Figure 2.8 (**A). Definition of epigenetic states used in this study. (B**)** The genomic space occupied by these states in the four cell types under study. (C**)** Median CpG content of the genomic regions in distinct epigenetic states. (D**)** Median expression level of epigenetic states used in this study based on assignment of each region to the nearest RefSeq gene. Median was computed over the states in all four-cell types and the corresponding expression profile. (E**)**. Size distribution of genomic regions enriched for at least one of our six histone modifications in at least one cell type and/or classified as UMR or IMR in at least one cell type. (F**)** Genomic features associated with all regions that change their epigenetic state in at least one cell type.

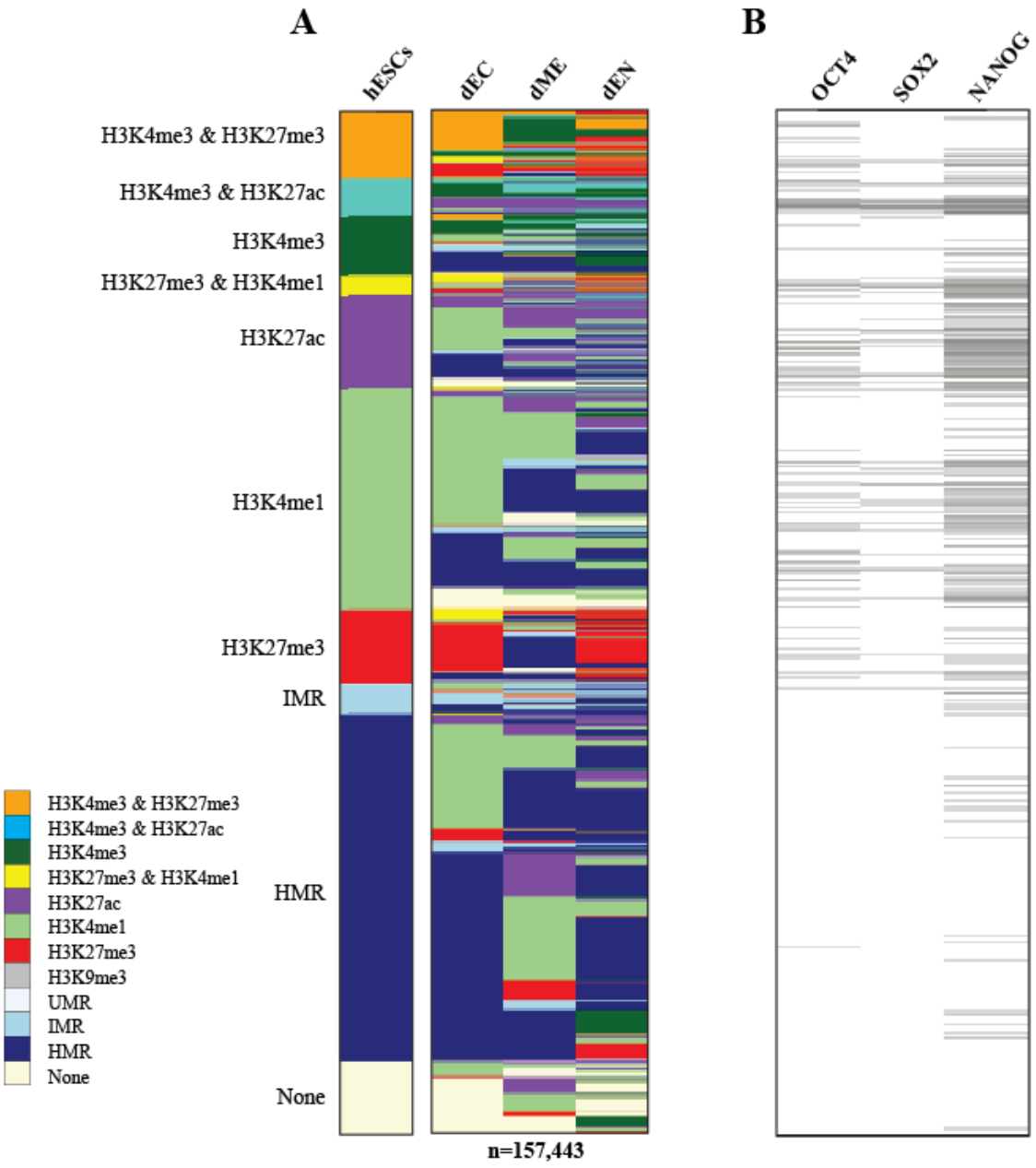



**Figure 2.9 (**A**)** Epigenetic state map of regions enriched for one of four histone modifications in at least one cell type or classified as UMR/IMR in at least one cell type and changing its epigenetic state upon differentiation in at least one cell type (n = 157,443). (B**)** Regions bound by OCT4, SOX2, and NANOG, as determined by ChIP-seq and organized using the chromatin states in A.

We also mapped O/S/N in HUES64, to identify epigenetic states commonly associated with their binding. All three factors were found at regions of euchromatin (**Figure 2.9B**), in particular those enriched for H3K4me1 and H3K27ac. The remodeling associated with their binding will be discussed in greater detail in Chapter 3.

Because previous reports suggested genes essential for development are held in a bivalent chromatin state in hESCs, we specifically explored the resolution of these domains during differentiation. We identified 4,639 proximal bivalent domains in hESCs and observed

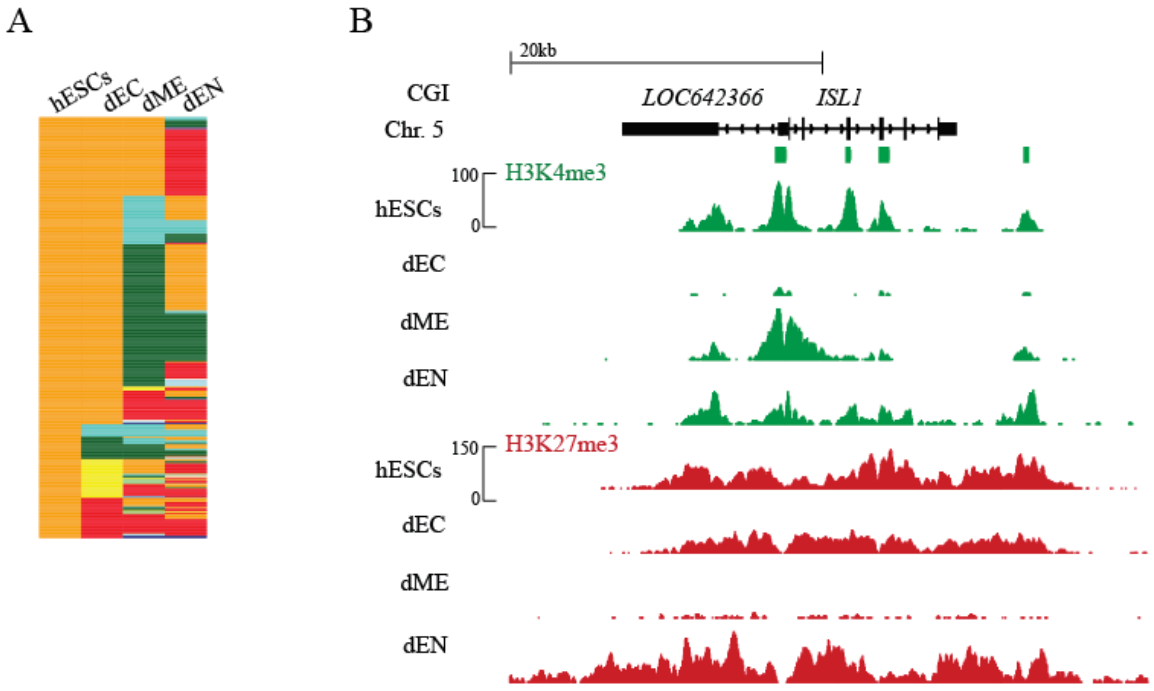



**Figure 2.10 (**A**)** Chromatin state map for all TFs that are bivalent in hESCs and change their epigenetic state in at least one cell type (n = 400). (B**)** Normalized ChIP-seq tracks of H3K4me3 and H3K27me3 at the *ISL1* locus (chr5: 50,661,163-50,703,879) indicating H3K27me3 is selectively maintained at high levels in dEC and dEN but not dME, where H3K4me3 increases while H3K27me3 is lost, promoting active transcription. Read counts on *y*-axis are normalized to 10 million reads and CGIs are indicated in green.

that 3,951 (85.1%) of these domains were no longer considered bivalent in at least one hESC-derived cell type (**Figure 2.9A**). Examining these domains more closely revealed that 463 of these domains are located at the promoters of TF-encoding genes, and 400 of these genes change their chromatin state in at least one differentiated cell type (**Figure 2.10A**). In dME and dEN, many transition to H3K4me3-only or H3K27me3-only in a lineage-specific manner, as shown for *ISL1* (**Figure 2.10B)**. In dME, H3K4me3 is gained at the *ISL1* locus while H3K27me3 is lost, leading to expression (FPKM: 14.3). The lineage specific dynamics in this region are interesting given that this gene has known roles in all three germ layers, although at later stages [145-147].

## 2.5 Conclusions and Discussion

The work presented in this chapter represents the development of a platform to observe epigenomic remodeling that accompanies lineage specification of hESCs. We first used NanoString profiling to monitor gene expression throughout five days of differentiation. This analysis showed that within 24 hours of differentiation induction, both the mesoderm and endoderm populations activated genes symbolic of the mesendoderm transition, such as *BRACHYURY*. After five days of differentiation, the NanoString profiles suggested that expression programs associated with early developmental stages of each germ layer were established. Expanding on our initial NanoString analysis using genome-wide expression signatures derived from RNA-Seq data, we found that each germ layer, as well as hESCs, exhibited few genes that were uniquely expressed. This result was most surprising for the dEC and dEN, because dEC differentiation involves inhibition of TGFβ signaling that is essential for endoderm specification. As with all of our results, we cannot exclude the possibility that heterogeneity existed within our populations, leading to biases within our analysis.

Using the expression level of the nearest transcript as a means to correlate epigenetic remodeling with expression, we also found that many remodeling events could not be linked to significant changes in gene expression. This result could be explained by our definition of association, which may not be sufficient because we did not consider long-range genomic interactions that are also associated with gene regulation [148, 149]. But given that most remodeling occurred at intergenic regions, the lack of significant changes in expression may have resulted because of differential enhancer usage. For example, an active gene may switch enhancer usage during differentiation, registering two remodeling events, though its expression may not significantly change. Two examples of this behavior were displayed at the *OCT4* and *NODAL* loci, which employ stage-specific enhancers during development [84, 150]. *NODAL* in particular had four nearby regions exhibit lineage specific remodeling.

The lack of correlation between epigenetic remodeling and changes in expression may have also resulted because of thresholds that we imposed, which could prevent identification of genes that remain expressed, but changed their expression level. If we look more closely at the expression level of genes expressed in multiple cell types, signatures associated with each population become more distinct. For example, *PAX6* is expressed at a higher level in the dEC (FPKM: 23.3) than in the dEN (FPKM: 5.21). The dEC and dEN may appear to be the most similar of the cell types that we interrogated because such information was disregarded in **Figure 2.5B**. Changes in gene expression and/or protein levels may change both the homo- and heterodimerization capabilities of proteins, which may in turn change a protein's affinity for DNA [117, 151].

Our splicing analysis suggested that both transcript levels and isoform expression also contribute to cellular identity, in addition to gene expression level. Alternative splicing of the

*DNMT3B* gene resulted in expression of the catalytically inactive isoform in each of the three germ layers, suggesting the translated protein may have assumed a new function in the differentiation process. This additionally suggests that a hallmark of pluripotency is expression of a *DNMT3B* gene that demonstrates de novo methylation activity.

The epigenomic profiling of each population also revealed distinct genome-wide epigenetic profiles at day five. 157,443 (≈58% of those identified) regions changed their epigenetic state during differentiation to at least one cell type. Most events were greater than 10kb from the nearest TSS, which correlated nicely with a recent report that found approximately 45% of DHS sites occurred at intergenic regions [152]. We also found that many bivalent domains are not retained during differentiation.

While induction of differentiation resulted in three unique epigenomic states at day 5, we specifically noticed an intriguing trend associated with H3K27ac enrichment. While regions that contained both H3K4me3 and H3K27ac exhibited the highest median expression level, not at all actively transcribed genes contained H3K27ac enrichment. In exploring this histone modification further, we found that H3K27ac enrichment occurred at genes that exhibit multiple transcription start sites, or genes that increased their level of expression in the differentiated cell types. For example, *DNMT3B* is down regulated during differentiation, exhibited an isoform expression switch and gained H3K27ac at the differentiation specific-TSS. Genes such as *TPM2, HAND1* and *DKK1* did not exhibit changes in isoform expression but gained this modification, and experienced greater than 2-fold upregulation during differentiation. We hypothesize that H3K4me3 enrichment is required for pre-initiation complex assembly [153], but that H3K27ac enrichment leads to accessible chromatin recognized by factors that can reinterpret transcript

information, potentially by inducing changes to Pol II pausing which was previously associated with alternative splicing [154].

In conclusion, we found that altering the signaling environment that maintains human pluripotency to mimic cues experienced during embryonic development causes genome-wide remodeling of the epigenetic landscape, with the majority of events occurring greater than 10kb from the nearest TSS. Our profiles identified many novel putative regulatory elements and the epigenetic mechanisms involved in their regulation, suggesting differentiation of hESCs is an excellent approach for studying human cellular transitions.

# Chapter 3.

# DNA Methylation Profiling Reveals Lineage-Specific Dynamics

The work discussed in this chapter is previously published. [123]

## 3.1 Rationale

While it is well established that DNA methylation is dynamic during differentiation and development, only recently have we begun to understand the discrete localization of specific regulatory elements that experience changes to DNA methylation. With the development of genome-scale sequencing assays, DNA methylation profiles of various populations were discerned. But many of these initial approaches included limitations that resulted in an incomplete appreciation of DNA methylation dynamics [155]. To overcome these limitations, the WGBS approach was developed. This assay provides the power to interrogate the methylation state of most cytosines within a genome in an unbiased manner, and has yielded important observations, such as the mutual exclusivity of transcription factor binding and high DNA methylation levels [10, 11]. We hypothesized that analysis of WGBS data collected from differentiated populations would reveal gene regulatory elements, and the associated genes, that are regulated by DNA methylation during lineage specification.

## 3.2 DNA Methylation Dynamics During Differentiation

Our WGBS libraries for each differentiated cell type, as well as HUES64, covered approximately 26 Million CpGs (at ≥5 coverage) across all four cell types. First, we performed hierarchical clustering analysis of the WGBS data to understand the genome-wide similarities between the populations, and included human adult liver and hippocampus for comparison. This revealed that the pluripotent hESCs and the hESC-derived cell types form a separate cluster arm with respect to the somatic tissues (**Figure 3.1**).

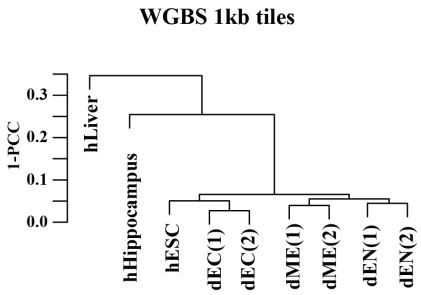



**Figure 3.1** Hierarchical clustering of hESCs, hESC-derived populations (dEC, dME, and dEN), human adult hippocampus, and human adult liver based on mean DNA methylation levels of 1 kb tiles across the human genome using Pearson Correlation Coefficient (PCC).

Next, we aimed to identify regions that change their DNA methylation state during differentiation at a statistically significant level. We defined differentially methylated regions (DMRs) as 1 kb windows that exhibited a CpG methylation level difference ≥0.1 between hESCs and the differentiated population ($p \leq 0.05$), using CpGs covered by ≥ 5 reads. We observed that the dEN exhibited far more regions that gained DNA methylation (n=1,963) compared to dEC (n=546) and dME (n=707) (**Figure 3.2A, top**), with the majority of regions found greater than 10kb in distance from the nearest TSS (**Figure 3.2A, bottom**). Most regions that increased their DNA methylation level were categorized as an IMR in hESCs, suggesting the switch from an unmethylated region to a highly methylated region was uncommon. Interestingly, only 65 of the total number of DMRs identified that gain DNA methylation were shared between all three populations (**Figure 3.2B**). However, reaffirming that our populations were depleted of pluripotent cells, this group of DMRs included the enhancer region of *OCT4 (***Appendix, Figure S5).** In examining these regions further, we found that most regions exhibited enrichment of one or more histone modifications in hESCs (**Figure 3.2C**). Promoters that gained DNA methylation

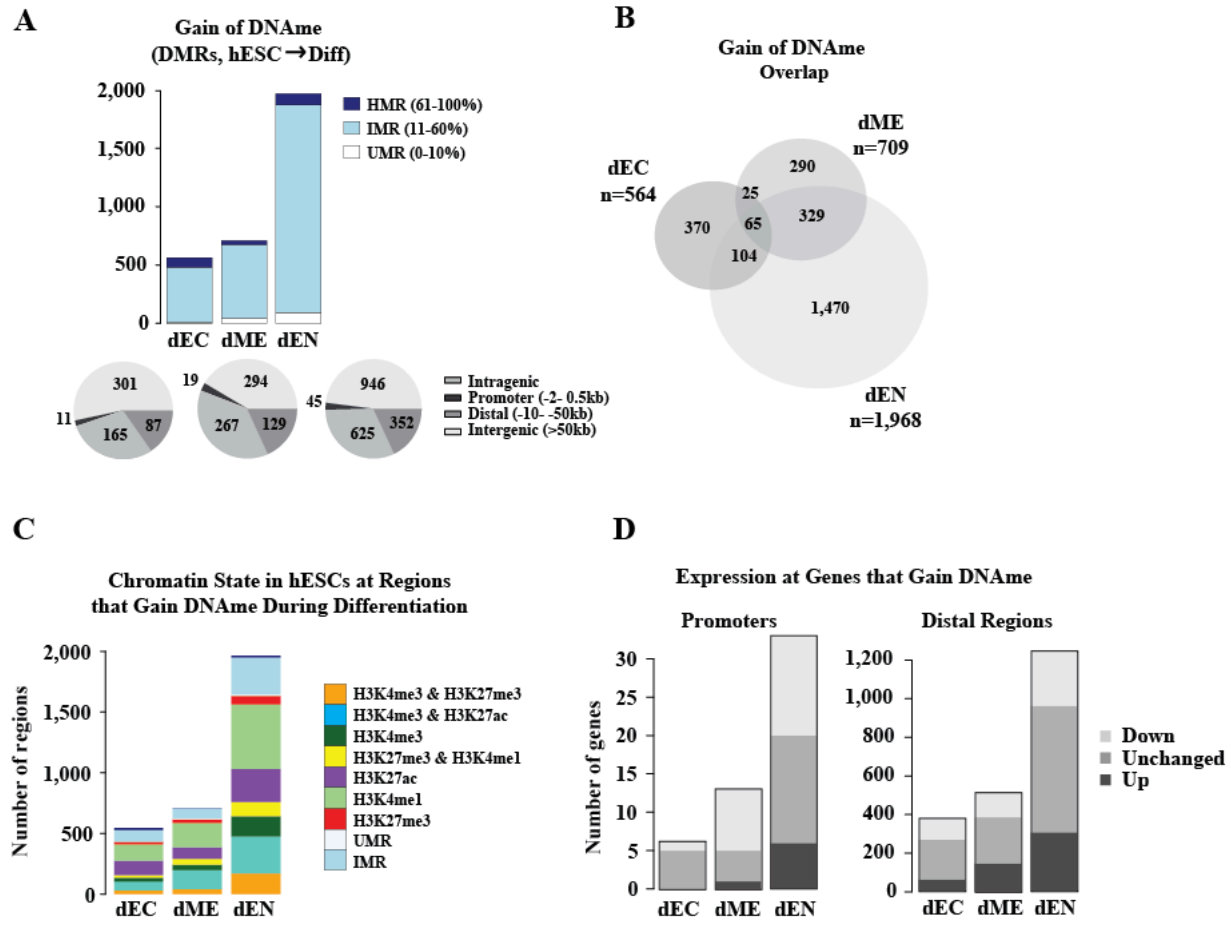



**Figure 3.2** (A) Top: Regions that significantly ($p \leq 0.05$) increase their DNA methylation levels by at least 0.1 between hESCs and the differentiated cell types. The color code indicates the DNA methylation state found in hESCs. Bottom: Genomic features associated with DMRs gaining DNA methylation in each of the differentiated cell types based on Ref Seq gene annotation and de novo discovered promoters by RNA-seq. (B) The overlap of differentially methylated regions (DMRs) that increase their DNA methylation level in the three hESC-derived populations. **(**C**)** Chromatin state in hESCs at regions that gain DNA methylation during differentiation. (D**)** Promoters (left) and distal elements (right) that gain DNA methylation separated by the changes in FPKM at associated genes.

in dEC and dME were associated with a decrease in expression (**Figure 3.2D**), while transcriptional silencing was less frequently correlated with gain of DNA methylation at distal elements (**Figure 3.2D**).

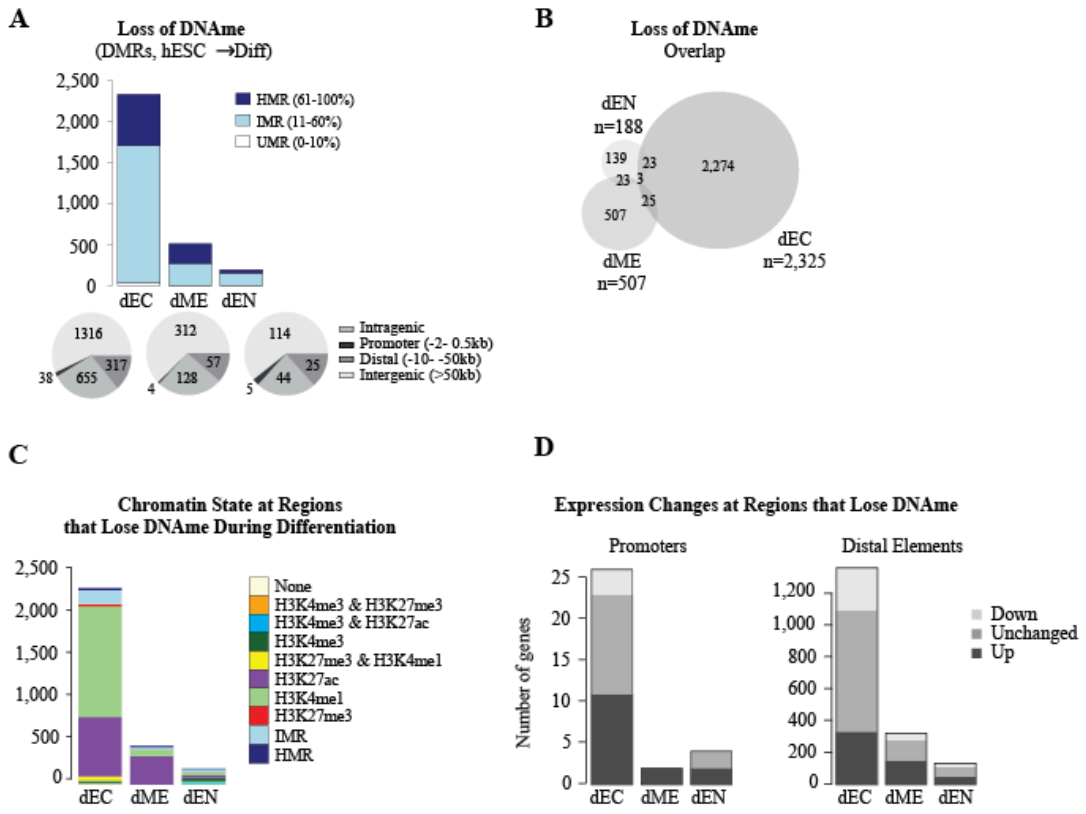



**Figure 3.3** (A) Regions that significantly ($p \leq 0.05$) decrease their DNA methylation levels by at least 0.1 between hESCs and the differentiated cell types. The color code indicates the DNA methylation state distribution in the differentiated cell types. Genomic features (bottom) associated with DMRs losing DNA methylation in each of the differentiated cell types based on Ref Seq gene annotation and de novo discovered promoters by RNA-Seq. (B) Venn diagram of identified DMRs that decrease their DNA methylation level between the three hESC-derived populations. (C) Chromatin state in differentiated cell types at regions that lose DNA methylation during differentiation. (D) Promoters (left) and distal elements (right) that gain DNA methylation separated by the changes in FPKM at associated genes.

Loss of DNA methylation was alternatively biased towards the dEC (n=2,313) **(Figure 3.3A**), but similarly occurred at regions that were not proximal to a TSS (**Figure 3.3A**). This transition occurred in a more lineage specific fashion than gain of DNA methylation, with only 3 regions shared by all three lineages (**Figure 3.3B**). More than 70% of DMRs that lose DNA methylation during differentiation were enriched for one of our profiled histone modifications in the differentiated populations, in particular H3K4me1 or H3K27ac (**Figure 3.3C**), marks

commonly found at enhancers. On a global scale, an immediate correspondence between loss of DNA methylation and expression occurred at approximately half of the regions (**Figure 3.3D**).

Regions that lost DNA methylation in the dEC were associated with neuronal gene categories (for instance: neural tube development, $q=3.13x10^{-13}$). This includes the ectodermal TF *POU3F1*, which had a bivalent promoter in hESCs, resolved to a H3K4me3-only state and exhibited transcriptional activation in dEC. Chromatin remodeling and activation at this locus coincided with specific loss of DNA methylation at a putative regulatory element downstream of the 3'UTR of this gene in dEC **(Figure 3.4)**.

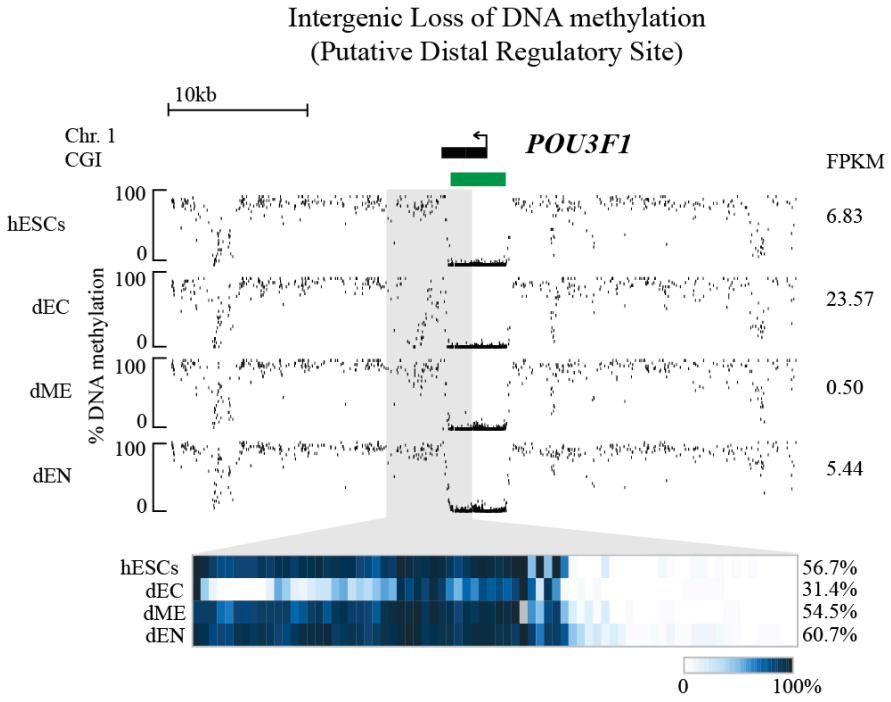



**Figure 3.4** DNA methylation at the *POU3F1* locus (chr1: 38,493,152-38,532,618). The heat map below shows the DNA methylation values of individual CpGs within the gray region. The average DNA methylation value for the entire highlighted region is shown on the right in red. CGIs are shown as green bars. Expression values (FPKM) are displayed on the right.

Examination of known regulators of signaling pathways involved specification revealed that the NODAL locus exhibits DNA methylation dynamics at multiple regions (**Figure 3.5.**).

The expression level of this gene is similar in both hESCs and the dEN (FPKM: 11.6, 10.3) in our system, while it is repressed in the dEC and dME (FPKM: 0.04, 0.14). NODAL employs two distinct cis-regulatory enhancers that control lineage and cell type-specific expression of the gene during pre-implantation development. Genetic studies in mice showed that deletion of the intragenic enhancer (ASE) limited *NODAL* expression to the proximal epiblast, while it was absent from the visceral endoderm and left lateral plate mesoderm [156, 157]. An upstream enhancer was alternatively necessary for expression in the node, and sufficient for induction of mesoderm [156, 158].

The ASE and the upstream enhancer gain DNA methylation in the dEC and dME, but not the dEN. The upstream enhancer is enriched for H3K4me1 in hESCs, and loses this mark while also gaining DNA methylation in the dEC and dME (**Figure 3.5**). Therefore, *NODAL* expression during definitive endoderm differentiation correlated with open chromatin at the ASE and upstream enhancer, suggesting these elements are also required during human post-implantation endoderm specification. We also identified an additional region further upstream that maintained H3K4me1 enrichment and low DNA methylation in hESCs, but it gained DNA methylation in all three hESC-derived populations. The gain of DNA methylation at the novel putative enhancer suggests that it is not required *NODAL* expression after specification of the germ layers.

## 3.3 Epigenetic Remodeling at OCT4/SOX2/NANOG Binding Sites

We next examined the DNA methylation state in the differentiated populations at regions bound by O/S/N in hESCs, factors included in the core pluripotency network [159], to understand if the previously reported depletion of DNA methylation in hESCs was maintained during differentiation [10].

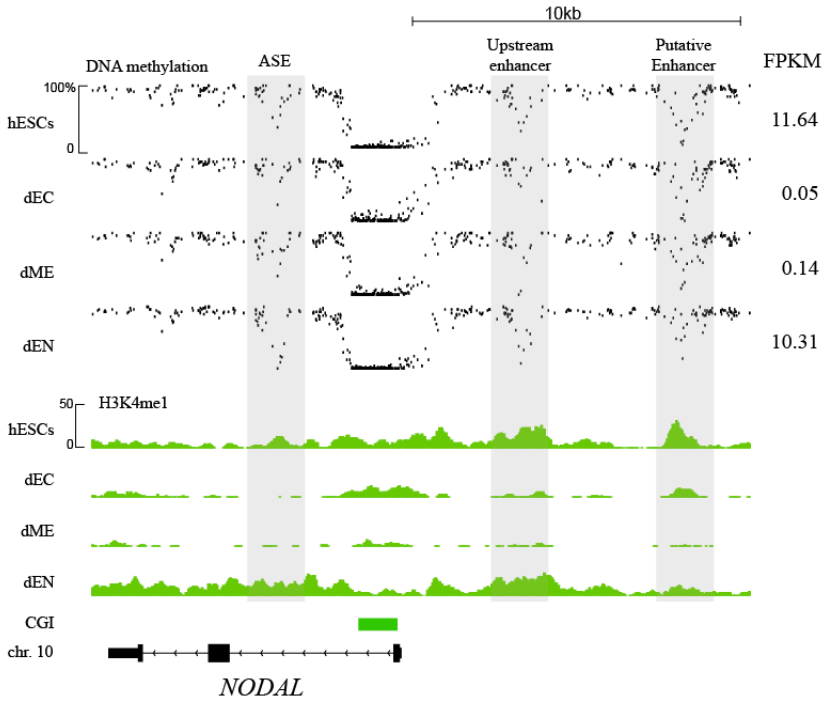



**Figure 3.5** WGBS and normalized H3K4me1 ChIP-Seq tracks for the *NODAL* locus (chr10: 72,191,148-72,213,104). Enhancers are highlighted in boxes.

NANOG exhibited the most unique binding sites (n=14,531), and SOX2 and OCT4 co-binding was the least frequent combination (n=956) (**Figure 3.6A**). 6% of identified sites were bound by all three factors, which was far less than previously reported [159]. Genome-wide, regions bound by all three factors (n=1,556), SOX2-only (n=923) or NANOG-only were frequently associated with intergenic regions rather than promoters (**Figure 3.6B**).

Epigenetic remodeling at regions of O/S/N binding were of particular interest given that these factors are reportedly involved in differentiation [130, 160], therefore we were curious as to fate of their binding sites. Our analysis revealed a lineage specific trend towards a heterochromatic HMR state (**Figure 3.7**). Examination of regions in hESCs bound by OCT4, NANOG and SOX2 individually showed H3K4me1 regions enriched for OCT4 binding sites frequently became HMRs in all three differentiated cell types, whereas NANOG and SOX2 sites were more prone

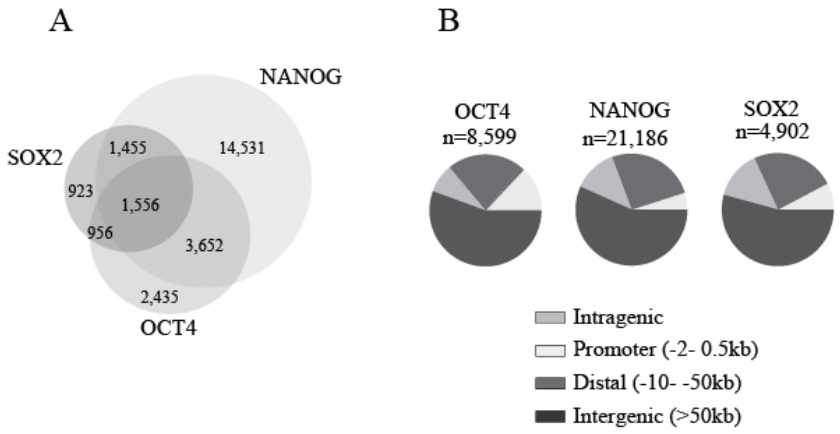



**Figure 3.6** (A) Venn diagram of the overlap between OCT4, NANOG and SOX2 binding sites identified in hESCs (total overlap = 1,556). (B) Genomic features of OCT4/NANOG/SOX2 binding sites.

to switch to an HMR state in dME (**Figure 3.7**). In general, many regions associated with open chromatin that were bound by NANOG, were more likely to retain this state in dEN compared to dME and dEC (**Figure 3.7**). We also found that regions enriched for H3K27ac in hESCs that maintain this enrichment in dEN or dEC, were likely to be bound by SOX2 and NANOG. This is in agreement with the reported role of SOX2 during ectoderm development and differentiation [161], but also supports our observation that SOX2 expression is maintained in the dEN. We additionally found a slight enrichment for NANOG binding at bivalent domains that transition to H3K4me3 only in the dEN. Furthermore, we found that regions bound by OCT4, NANOG and SOX2 that gained an active mark in dEC were enriched for the *PAX9* motif, a gene associated with later stages of neural crest development [162]. Inspection of DMRs associated with O/S/N binding sites revealed that many genes bound by these factors were associated with later stages of development rather than pluripotency, and they exhibited lineage specific remodeling dynamics. For example, two regions 20 kb downstream of *DBX1*, a gene associated

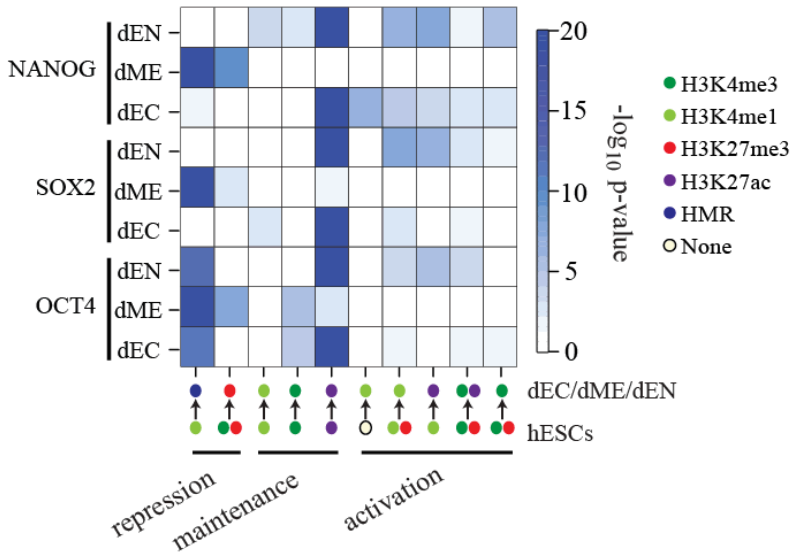



**Figure 3.7** Enrichment of OCT4, SOX2, and NANOG within various classes of dynamic genomic regions changing upon differentiation of hESC, computed relative to all regions exhibiting the particular epigenetic state change in other cell types. Epigenetic dynamics are categorized into three major classes: repression (loss of H3K4me3 or H3K4me1 and acquisition of H3K27me3 or DNA methylation), maintenance of open chromatin marks (H3K4me3, H3K4me1, and H3K27ac), and activation of previously repressed states.

with neural specification, were bound by all three TFs in hESCs and they gained DNA methylation in dME and dEN (**Figure 3.8**). However, they maintained low levels of DNA methylation in the dEC, the same population where active transcription was also detected (**Figure 3.8**). This may also explain why distal regions that gained DNA methylation did not exhibit a decrease in expression, as many regions may have been held in an open state for activation at a later time.

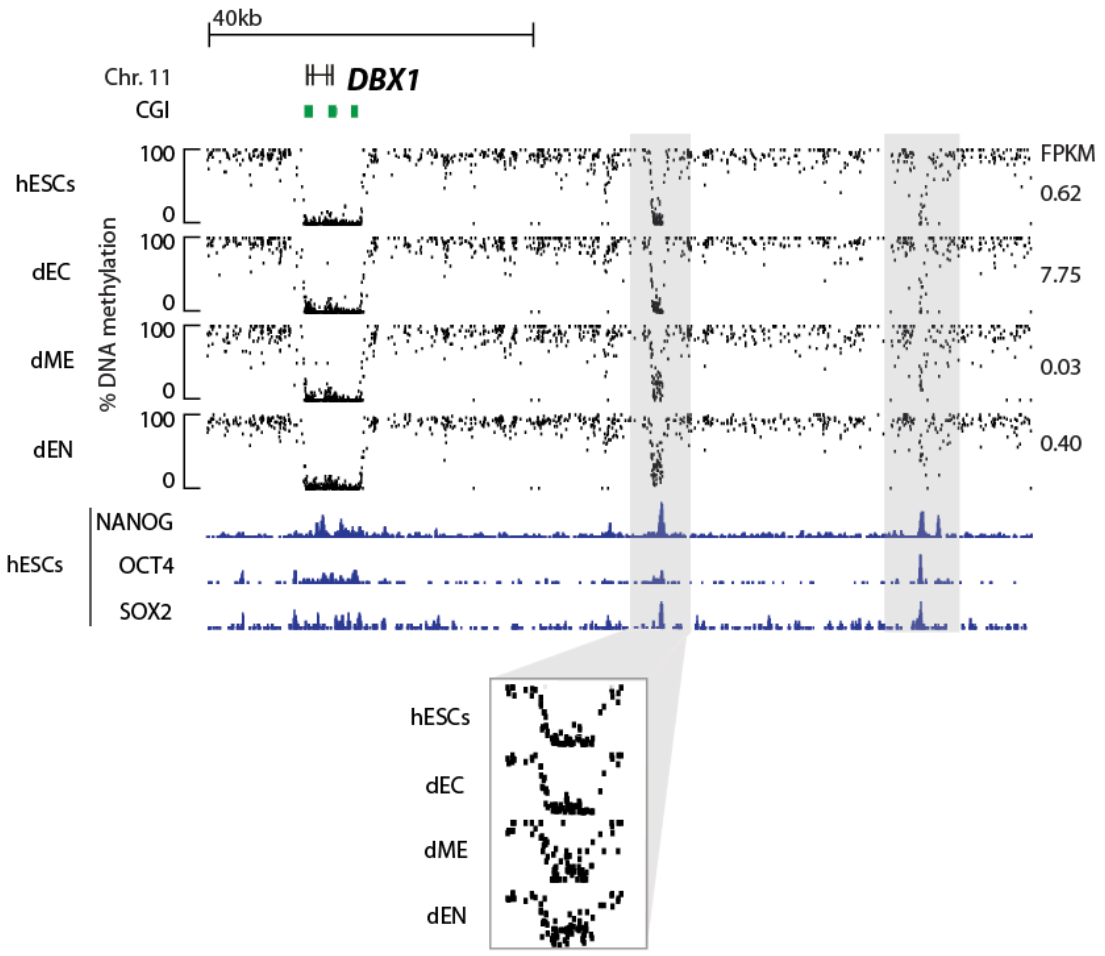



**Figure 3.8** DNA methylation levels and OCT4, SOX2, and NANOG ChIP-seq at the *DBX1* locus (chr11: 20,169,548-20,277,940).

## 3.4 Conclusions and Discussion

Taken together, this chapter prevents evidence that DNA methylation is quite dynamic during differentiation and exhibits an intriguing lineage bias. In contrast to a previous study, we found that changes in DNA methylation status are not commonly correlated with altered gene expression [11]. The dEN displayed the most regions that gain DNA methylation, and they were more commonly shared between differentiated cell types (n= 520, 20% of DMRs). This is likely because gain of DNA methylation represents lineage specific silencing of regulatory regions associated with pluripotency, an essential task each lineage must perform. We also found

significant enrichment of various TF motifs associated with alternative lineages that are DNA methylation targets upon differentiation**,** which has some analogy to the gain of methylation observed at myeloid targets in the lymphoid lineage *in vivo* [16]. It is thought that regulatory elements required for all lineages are maintained in a euchromatic environment in the pluripotent state to promote ease of activation downstream, therefore once a lineage is a specified, these elements can be stably silenced by DNA methylation as they likely will not be required later in development.

The bias towards gain of DNA methylation in the endoderm may result because NODAL/TGFβ signaling plays a prominent role in both promoting human pluripotency and directing endoderm specification, but they may employ different enhancer elements. Similar to the mechanisms of *OCT4* [150] and *NODAL* regulation [156], DNA methylation may distinguish distal regulatory elements available for regulatory purposes for each state. The gain of DNA methylation may reflect a lineage choice that requires only a specific branch of TGFβ signaling. This concept of gradual restriction is a central theme in embryonic development.

The asymmetric loss of DNA methylation was the most prevalent in the dEC and was highly lineage specific, with only 3 regions losing DNA methylation in all three lineages. Though each lineage exhibited gain of the enhancer-related histone modifications H3K4me1 and H3K27ac. In cooperation with our distance-to-TSS analysis for regions that loose DNA methylation, this supports recent evidence that enhancer elements remodel prior to promoters of associated genes [100]. We hypothesize that these events represent the initial stages of activation of enhancer regulatory elements required for ectodermal-differentiation that are not held in a euchromatic state in hESCs.

Examining epigenetic dynamics at sites of O/S/N binding reveals that regions bound by one of these factors also showed a lineage specific preference for maintaining open chromatin in the differentiated cell types. Many regions bound by these factors were also not expressed in hESCs, but were associated with genes required at later stages of development, such as *DBX1*. Two discrete regions associated with this gene gained DNA methylation in the dME and the dEN, but not the population in which it was activated (dEC). This data suggests that O/S/N may serve an additional purpose by maintaining an open chromatin environment that allows exploitation of the regulatory element at a later time point. Additional support for their role in chromatin structure was presented by a report that found these factors are commonly bound with mediator, a protein essential for maintaining chromatin structure and interchromosomal looping [163]. This report was recently expanded on by the identification of "super-enhancers," which found that multiple master regulators of cell fate, including Oct4 and Mediator, are commonly bound in close proximity to one another. These regions of concentrated binding were essential for maintaining the pluripotent state *in vitro*, but the dependency on the individual factors for proper chromatin architecture is unresolved.

While studies relating Mediator to O/S/N binding associate these factors with chromatin structure, their ability to induce remodeling of chromatin versus maintaining previously established states is currently somewhat controversial. In an attempt to understand the chromatin binding affinity and remodeling capabilities of the reprogramming factors OSKM, one group mapped the binding sites of these proteins in adult fibroblasts at 48 hours –post OSKM induction. They found that most binding sites were distal to the TSS and they frequently overlapped regions resistant to DNaseI, a sign of closed chromatin, or H3K9me3. This led to their conclusion that OSKM can act as pioneers at genomic regions of closed chromatin. This conclusion may have

been premature, as less than 50% of regions exhibiting 5hmC in hESCs overlapped with DHS sites [164]. Therefore, it is unclear what degree of heterochromatin is exhibited at OSKM bound sites that are resistant to DNaseI. Given that proteins exhibit distinct binding affinities depending on the methylcytosine state [165], we suggest that OSKM should not be considered traditional pioneer factors using their data set. However, an independent study identified 19 proteins present in mESC nuclear extracts that preferentially bound 5mC rather than unmethylated cytosine, including Klf4 [165].

The mechanism by which OCT4, or any TF, is eliminated from its binding sites during differentiation and development to allow or promote repression is also unresolved. Eviction of OCT4 from its binding sites may be induced directly by gain of DNA methylation. Alternatively, POU factors do not remain bound to mitotic chromosomes, therefore addition of DNA methylation may prevent OCT4 from interacting with its binding sites after cell division is complete [120]. If this scenario were true, it suggests that additional factors maintain NDRs at OCT4 binding sites during cell division. While the evidence was limited, one report suggested an analogous mechanism for FoxD3 in maintaining a DNA methylation-free CpG upstream of the Alb gene, which promoted its efficient activation at a later stage of differentiation [166]. This phenomenon, also known as bookmarking, has been suggested for TFs such as GATA1 and FOXA1, which remain bound to mitotic chromosomes [122, 166]. While these two factors are not expressed in hESCs, many similar proteins such as the aforementioned FOXD3, as well as FOXB1 and FOXI3, are actively expressed in hESCs.

In conclusion, this chapter includes evidence that DNA methylation is quite dynamic intergenic regions bound by TFs during lineage specification, though changes are discrete and localized. We confirmed that regulatory elements associated with pluripotency gained DNA

methylation in each lineage, suggesting they were stably repressed. A lineage bias was also detected within the DNA methylation analysis that may have interesting implications for understanding the regulatory networks employed by each lineage, as well as hESCs. Our O/S/N binding data also suggests that these TFs may be involved in maintaining an open chromatin state at regulatory elements that are required during later stages of development, rather than just in the pluripotent state.

# Chapter 4.

# Activation of Somatic-related Regulatory Elements Through Epigenetic Priming

The work presented in this chapter is previously published. [123]

## 4.1 Rationale

While many reports have studied the dynamics that lead to silencing the pluripotent network, few studies have elaborated on the activation of lineage specific programs. The events that cause loss of high DNA methylation levels at genes required during later stages of specification, creating an environment amenable to transcriptional activation, is not well understood. These events are essential to the specification process as many genes associated with somatic cell types exhibit high DNA methylation at their regulatory elements in ESCs [14, 167].

## 4.2 Acquisition of H3K4me1 at hESC-HMRs

To identify genes that transitioned to a euchromatic state during differentiation, we examined the epigenetic states of genes associated with downstream stages of development and differentiation. We found many regions that exhibit high DNA methylation in hESCs, and transition to H3K4me1 in one lineage, but remain highly methylated in the two alternative cell types (**Figures 2.10A and 4.1A**), analogous to remodeling events reported during reprogramming and cardiac differentiation [168, 169]. This remodeling trend occurred in all three lineages, providing evidence that this is a common transitory occurrence. Similar to the regions showing dynamic DNA methylation during differentiation, these regions were typically intergenic (**Figure 4.1B**).

To understand if regions that gain H3K4me1 are associated with somatic identity, we took advantage of published microarray data for 24 human tissues and determined genes upregulated in these tissues with respect to hESCs. Reaffirming the relevance of our dynamics, we found regions that gain H3K4me1 in dEC were associated with fetal brain, cortex and cerebellum (**Figure 4.1C**) based on region association with the nearest gene. The dME

H3K4me1 pattern was associated with the heart, as well as unexpected tissues, which may be due to heterogeneity of the tissues collected. The dEN was associated with lung and pancreas, as well as fetal brain and cortex (**Figure 4.1C**).

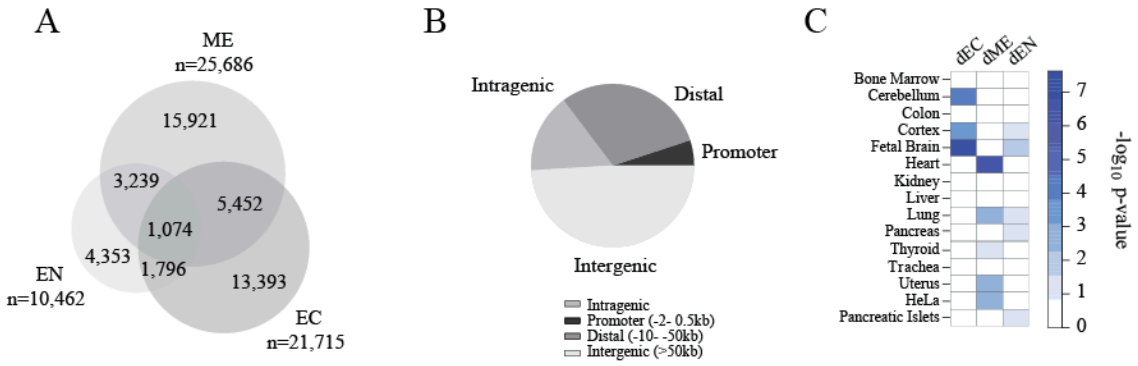



**Figure 4.1** (A) Overlap of regions gaining H3K4me1 in the three differentiated populations relative to hESCs. (B) Genomic distribution of all regions gaining H3K4me1 compared to hESCs in at least one of the three differentiated populations. (C) Tissue signature enrichment levels of genes assigned to regions specifically gaining H3K4me1 in the differentiated populations indicated on the bottom.

To investigate these regions in more detail, we carried out motif enrichment analysis and found lineage specific enrichment of TF motifs near regions that gain H3K4me1. While the FOXA2 motif was enriched in all three cell types, the DBX1 motif was associated with the gain of H3K4me1 in dEC (**Appendix, Figure S7**), which coincided with its transcriptional activation in that cell type (FPKM: 5.36). Conversely, the GLI3, HIC1 and CTF1 motifs were strongly enriched at regions that gain H3K4me1 in dEN (**Appendix, Figure S6).** Overall, less than half of the genes that gain H3K4me1 exhibited immediate transcriptional changes (**Figure 4.2A**). *CYP2A6* and *CYP2A7* (**Figure 4.2B**) are representative examples that did not show a corresponding change in expression, while *LMO2* was activated in the dME (**Figure 4.2C**).

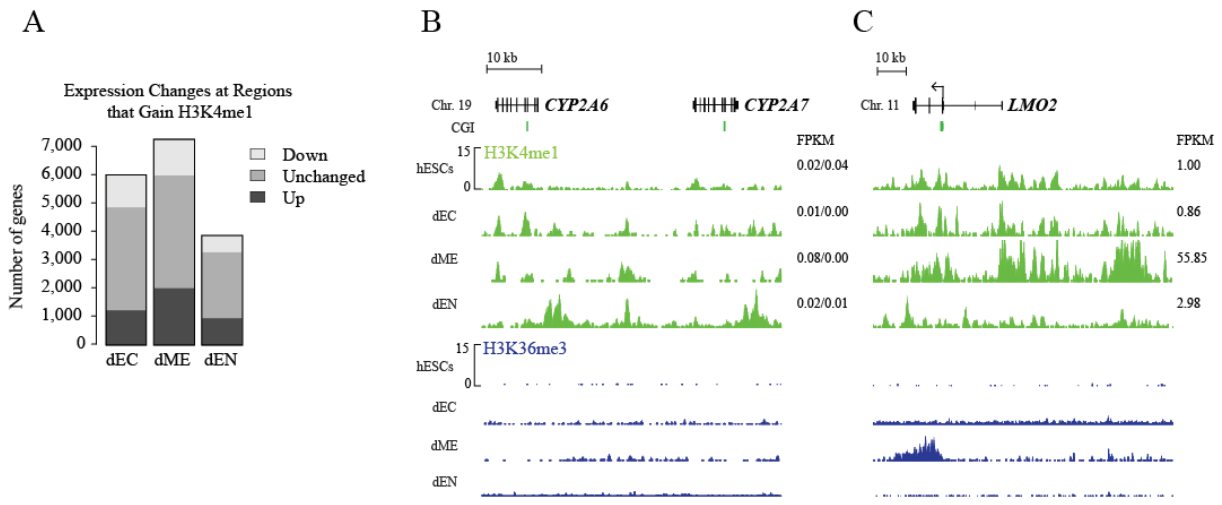



**Figure 4.2** (A) Number and distribution of gene expression changes of genes assigned to regions gaining H3K4me1 in the differentiated populations. Associated genes were classified as either being up/downregulated or unchanged relative to hESCs. (B) Normalized ChIP-seq tracks (H3K4me1 and H3K36me3) for the *LMO2* locus (Chr.11: 33,865,134-33,977,858). Read counts on *y*-axis are normalized to 10 million reads for each cell type. CGIs are indicated in green. (C) Normalized ChIP-seq tracks (H3K4me1 and H3K36me3) for the *CYP2A6/CYP2A7* region (Chr19: 41,347,260-41,395,599). Read counts on *y*-axis are normalized to 10 million reads for each cell type. CGIs are indicated in green.

To further assess if this DNA methylation-to-H3K4me1 switch acts as an epigenetic priming event, we differentiated the HUES64 endoderm population for five additional days in the presence of BMP4 and FGF2, leading to a hepatoblast-like (dHep) state and then performed RNA-Seq [132]. Interestingly, motifs enriched in dEN that gained H3K4me1, including HIC1 (FPKM 1.46) and CTF1 (FPKM 5.32), became expressed at the next stage of differentiation, along with markers of the corresponding stage of hepatocyte differentiation, such as *AFP* (FPKM: 5.16) (**Figure 4.3**), and *HHEX (*FPKM: 3.32*)*.

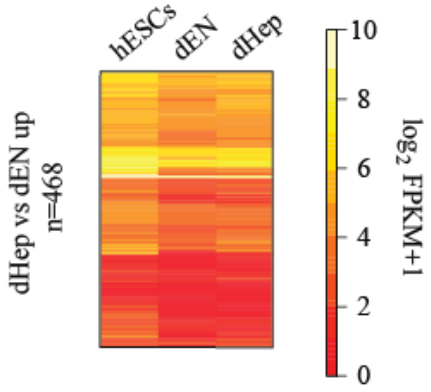



**Figure 4.3** Gene expression levels of genes being upregulated between dEN and dHep (but not between hESC and dEN) and gaining H3K4me1 in dEN are shown.

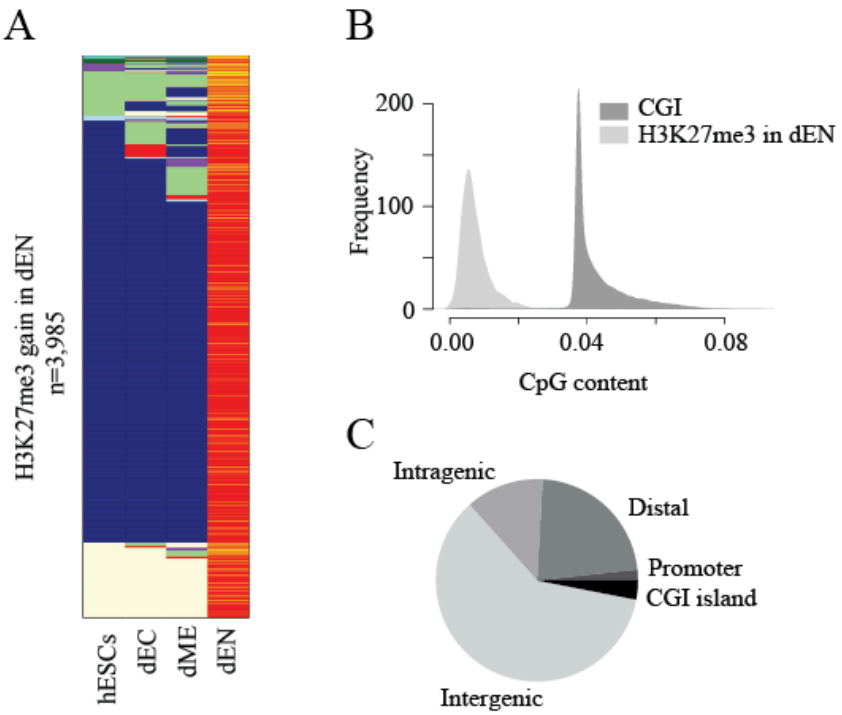



**Figure 4.4** (A) Epigenetic state distribution in hESC, dEC, and dME of regions that gain H3K27me3 in the dEN population compared to hESC. (B) CpG content distribution of regions gaining H3K27me3 upon differentiation. For reference, the CpG content distribution of CpG islands is shown. (C) Distribution of genomic features associated with regions gaining H3K27me3 (n = 22,643).

4.3 FOXA2 Binding is Associated with Epigenetic Priming

We also observed that many intergenic regions switched from high DNA methylation to H3K27me3 in a lineage specific manner (n=3,985 in dEN) (**Figure 4.4A**) while maintaining the HMR state in the alternative cell types. This transition frequently occurred within CpG poor intergenic regions (**Figure 4.4 B-C**), which is distinct from the common CpG island-centric targets of PRC2.

In an effort to identify factors that may be directing this remodeling, we utilized TF binding data from the ENCODE project [170]. By overlaying our epigenetic profiles with their TF binding data, we found many regions that exhibited the transition from HMR to H3K27me3 enrichment in CpG poor regions in the dEN were near FOXA2 binding sites identified in the HepG2 cell line. FOXA2 is essential for endoderm development [171, 172], and *in vitro* studies showed that it binds more stably to nucleosomes rather than free DNA [173], which leads to chromatin decompaction *in vitro* [107]. Although, due to technical limitations associated with the *in vitro* assays employed in the aforementioned studies, FOXA2's distinct chromatin remodeling-related functions and limitations remain somewhat unclear.

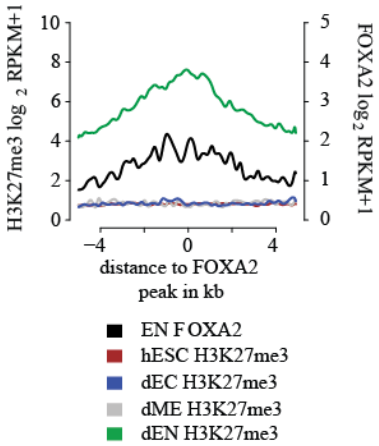



**Figure 4.5** Composite plot of median normalized tag counts (reads per million RPKM) of regions bound by FOXA2 in dEN and gaining H3K27me3 in dEN compared to hESC (n = 357).

We therefore hypothesized that this "pioneering" TF may induce this epigenetic transition from high DNA methylation to H3K27me3 enrichment. To investigate this association, we performed ChIP-Seq for FOXA2 in the endoderm population. As shown by the composite plot, this analysis revealed that FOXA2 binding sites frequently overlapped with regions that transition from a HMR state to a H3K27me3-containing state (**Figure 4.5**). In addition, this gain of H3K27me3 at FOXA2 binding sites predominantly occurred in dEN (**Figure 4.5)**. A notable example of this transition was seen at the albumin locus, where H3K27me3 was gained at *AFP* and *AFM*, proximal to FOXA2 binding sites **(Figure 4.6).** This mark was not found at this locus in primary liver tissue, potentially suggesting it represents a transient state (**Figure 4.6**).

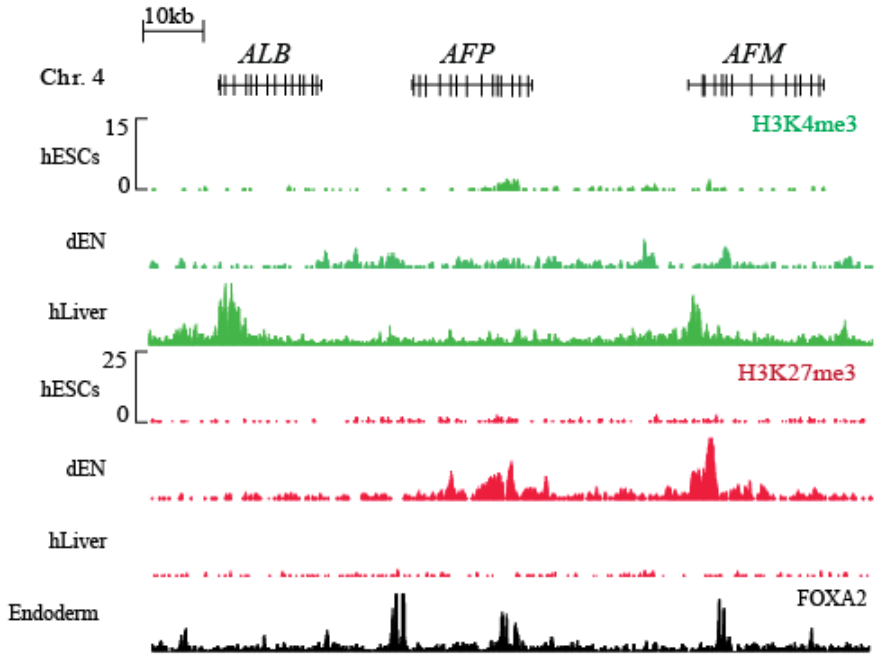



**Figure 4.6** Normalized H3K27me3 and H3K4me3 ChIP-seq tracks for hESCs, dEN, and human adult liver tissue at the *ALB* locus (chr4: 74,257,882-74,377,753).

Many regions that exhibit this transition are required for later stages of development as with *AFP*, or the hemoglobin locus in the dME

Because H3K27me3 was gained at hESC-HMRS, next we compared DNA methylation at FOXA2 binding sites in hESCs to dEN, and found a slight reduction in the dEN **(Figure 4.7).** To more directly assess this relationship, we then interrogated the DNA methylation state of regions isolated by FOXA2-ChIP in the endoderm, also known as ChIP-bisulfite-sequencing (ChIP-BS-Seq) [174]. This direct assay revealed a major depletion of DNA methylation at sites isolated by FOXA2-ChIP (**Figure 4.7**). The methylation value appeared only slightly depleted in the WGBS data because we computed the methylation value using each CpG under the FOXA2 peak, which included far more CpGs than those covered by ChIP-BS-Seq.

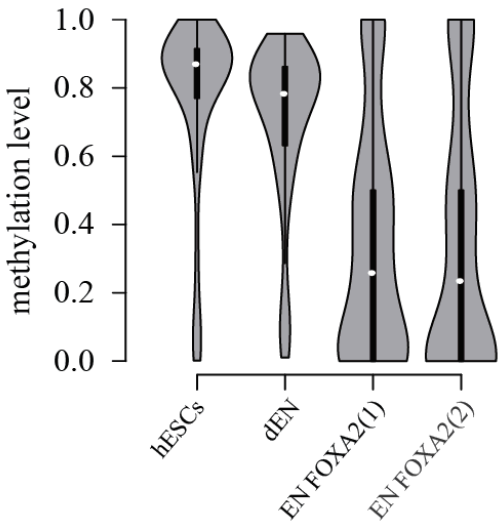



**Figure 4.7** Distribution of methylation levels of regions bound by FOXA2 and gaining H3K27me3 in dEN. DNA methylation information is depicted for hESC and dEN WGBS data sets and two biological replicates of FOXA2 ChIP-bisulfite experiments in dEN (n = 357).

To determine if these regions exhibited transcriptional activation after further differentiation, we again examined our dHep RNA-Seq data. We found that 50 genes were bound by FOXA2, gained H3K27me3 in dEN and increased their expression in the dHep population

(**Figure 4.8**), including *AFP*. Analysis of H3K27ac ChIP from human liver also revealed enrichment of 197 loci that had experienced gain of H3K27me3 in dEN, suggesting that these loci eventually transitioned to an active state (**Appendix, Figure S7**). We believe that de novo gain of both H3K27me3, as well as H3K4me1, serve as priming events that create an epigenetic state amenable to binding of additional factors that are not capable of inducing epigenetic remodeling, leading to gene activation at later stages of differentiation [100].

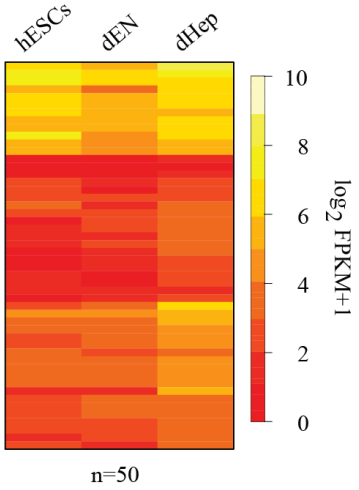



**Figure 4.8** Gene expression profile of genes upregulated at the hepatoblast stage relative to dEN that are associated with regions bound by FOXA2 and gaining H3K27me3 in dEN (n = 50)

## 4.4 Conclusions and Discussion

The evidence summarized in this chapter suggests that the transition from high DNA methylation to a euchromatic state exhibiting enrichment of histone modifications frequently involves a transient state that we believe is a novel "primed" epigenetic state. We frequently detected HMRs that gained enrichment of H3K4me1 or H3K27me3, but they were not necessarily associated with transcriptional activation. We confirmed that a subset of related genes discovered in the dEN experienced activation during differentiation to the next stage towards hepatocytes. Recent studies have reported dynamics that suggest similar epigenetic

priming events occur during cardiac differentiation [168]. These results are also reminiscent of changes that occur during the early stages of reprogramming towards the induced pluripotent state and highlight possible similarities between differentiation and de-differentiation. By isolating populations based on the number cell divisions post-OSKM induction, our lab previously reported that initial gene expression dynamics were predominantly observed at loci containing a preexisting euchromatic state as defined by H3K4 methylation and/or K3K27 methylation [175]. In that study, a de novo or enhanced gain of H3K4me2 was observed at promoters of many pluripotency-associated genes prior to detectable transcription originating from these loci.

In an effort to identify individual transcription factors that are associated with these priming dynamics, we examined previously published TF binding data and observed that pioneer transcription factor binding was associated with regions that gain H3K27me3 during hESC-differentiation. FOXA2 binding profiles generated from the endoderm population confirmed its correlation with this epigenetic remodeling event. We additionally confirmed by ChIP-BS-Seq that these regions exhibited loss of DNA methylation during differentiation.

Further analysis of the ChIP-BS-Seq was confounded by additional factors. DNA fragment sizes ranged from 300-600bps within our sequencing library, but TF motifs/binding sites range in size from 8-20bp. Consequently, we cannot distinguish between the two following scenarios; FOXA2 directly binds to methylated cytosines, or FOXA2 binds in the proximity of methylated cytosines. In the dEN, loss of DNA methylation was confined to FOXA2 peaks, but these peaks were greater than 400bp. Therefore, while we cannot conclude if FOXA2 binds methylated DNA directly, it appears that FOXA2 does not exert remodeling capabilities that spread throughout genomic regions. While it has previously been suggested that FOXA1, a

paralog of FOXA2, has the ability to bind methylated regions and induce local remodeling, this function appears to be limited to a subset of TFs [1, 165]. Among many other interesting observations, Stadler et al also showed that similar to our observations regarding FOXA2, CTCF binding appeared to induce local demethylation, within 500bp of the binding site. Extracting more discrete information from this data was further complicated by the fact that FOXA2 bound regions were not CpG dense, and therefore encoded approximately 1 CpG/100bp. We only collected 50bp of sequence, severely limiting the number of reads that contained multiple CpGs. Addition of 5hmC enrichment may dramatically enhance our ability to address these lingering questions [164].

It is also unclear whether binding recruits additional factors necessary for remodeling, or whether FOXA2 induces remodeling independently by inhibiting/stimulating epigenetic modifiers that it naturally interacts with while it's bound to DNA. *In vitro* studies showed that FOXA factors could induce opening of chromatin independent of ATP, suggesting pioneer factor binding itself may be sufficient to initiate opening, though the DNA template did not include methylcytosine. FOXA2 peaks additionally appeared quite broad in the composite plot, though closer investigation revealed that multiple individual peaks were often identified close together (within 1kb), including at regions that gained H3K27me3. An *in vitro* system, which utilized the ALB enhancer sequence, found that FOXA2 binding directly altered nucleosome placement [176] can decompact a chromatin array *in vitro* [177]. With multiple binding sites close together, nucleosome placement and decompaction is potentially more finely controlled. Enrichment of H3K27me3 during NPC differentiation was also elevated when multiple SNAIL binding sites occurred in close proximity [83].

Given that FOXA2 binding did not result in transcriptional activation of many loci in the endoderm population, this provides additional evidence that FOXA2 binding may be the first step in a cascade of events, which paves the way for additional factors to bind and promote transcriptional activation at later stages. Future studies should be directed at understanding the sequence of events, and establishing that these events are necessary for the timely activation of gene expression during specification.

In conclusion, this chapter provides evidence that the process of activating regions required for differentiation involves the loss of DNA methylation and/or gain of histone modifications, such as H3K4me1 and H3K27me3. By defining discrete regions that undergo these transitions, we have also putatively defined factors that direct these events. These epigenetic transitions are not always coupled to significant changes in expression, which we hypothesize represents epigenetic priming that will allow activation of these loci at later stages of differentiation.

# Chapter 5.

# Discussion and Future Studies.

## 5.1 Summary

Dynamic epigenetic regulation has been observed during the acquisition of new cellular identities, such as during embryonic development. While basic associations with transcriptional activity have previously been ascribed to epigenetic mechanisms, the precise regulatory elements that DNA methylation and histone modifications control to facilitate differentiation have not been widely studied in a human context for technical and ethical reasons. Thus, we chose to use 2-dimensional directed differentiation of HUES64, a male hESC line, towards each major embryonic lineage to capture epigenetic remodeling that accompanies *in vitro* specification (**Figure 2.1**). During differentiation, each population exhibited transcriptional dynamics similar to those previously observed *in vivo*, such as the induction of *BRACHYURY* during mesoderm and endoderm differentiation (**Figure 2.2**). We employed FACS-based isolation to increase the homogeneity of the day five populations, and subsequent expression profiling revealed that each population resembled one of the three embryonic germ layers (**Figure 2.4**). Herein we will refer to FACS sorted populations as dEC (ectoderm), dME (mesoderm) and dEN (endoderm). Surprisingly, the global transcriptional landscape defined by RNA-Seq was quite similar between the three differentiated cell types and the hESCs (**Figure 2.5**), though notable differences between absolute expression levels and splicing were observed (**Figure 2.6**).

Next, we performed ChIP-Seq for six histone modifications (H3K4me1, H3K4me3, H3K27me3, H3K27ac, H3K36me3 and H3K9me3) and WGBS to determine DNA methylation levels, on the three derived populations as well as HUES64. Integration of these data sets revealed that each population displays a distinct epigenomic state, categorized by widespread remodeling of both histone modifications and changes to DNA methylation levels compared to the undifferentiated HUES64 (**Figure 2.9**). We found expected commonalities between lineages,

such as the stable repression of regulatory elements associated with the pluripotent network (**Figure S6**) and alternative lineages (**Figure 3.8**). Alternatively, we also saw that enrichment of euchromatic histone modifications at O/S/N binding sites was sometimes retained in the differentiated cell types, suggesting a subset of regions bound by O/S/N may be involved in later stages of differentiation (**Figures 3.7** and **3.8**). We also uncovered regions of facultative heterochromatin that exhibit lineage specific loss of DNA methylation and/or gain of histone modifications such as H3K4me1 or H3K27me3 during differentiation (**Figures 2.9, 4.1** and **4.4**). This transition was pronounced at genes that did not concurrently exhibit transcriptional activation but are associated with later stages of differentiation in their respective lineages (**Figures 4.3 and 4.8**). We therefore termed these regions "primed," to distinguish them from the common pool of repressed genes. While we cannot confirm that these events occur during embryonic development because we did not probe primary tissue, similar observations have been made *in vivo [168]*.

We also observed general trends throughout our analysis, such frequent epigenetic remodeling events that we cannot associate with significant changes in transcription. This is potentially explained by the observation that most events are intergenic, occurring greater than 10kb from the nearest TSS, suggesting the majority of remodeling occurs at enhancer elements. A major lineage bias was also observed in the WGBS data, in that the dEC gained DNA methylation at many more regions than the dME and dEN, while the dEN had the most regions that lost appreciable amounts of DNA methylation. Using epigenetic dynamics as a proxy for regulatory significance, our data represents an expansive resource for identifying novel regulators of human embryonic specification.

## 5.2 Transcriptional Signatures Reveal Few Differences

Based on RNA-Seq, our three differentiated cell types, as well as the hESCs, are quite similar on a gene expression level (**Figure 2.5**). For our analysis, we used annotations composed of mostly protein-coding elements with limited non-coding RNAs, suggesting that our similarities are due to protein-coding genes, which may be expressed ubiquitously throughout development. This is similar to a claim made by the ENCODE consortium where RNA sequencing of 14 cell lines that represented various lineages suggested that 53% of protein-coding genes were detected in all cell lines, while only 7% were unique to one cell type [178].

The sensitivity of the techniques that we used to measure expression afforded us the ability to make many interesting observations regarding expression. NanoString and RNA-seq generated quantitative, digital expression values that did not rely on comparison to a starting cell type for normalization. The limiting nature of alternative approaches, such as PCR and microarrays, may explain why SOX2 is not widely associated with definitive endoderm differentiation. This gene is downregulated during mesoderm and endoderm differentiation in our system, but quantitative approaches employed here allowed the observation that transcripts are still present in the dEN, suggesting the gene is not yet repressed. Growing evidence suggests that SOX2 is essential for endoderm development [179] and organ function [180].

The NanoString and RNA-Seq analyses also reveal that to define the current lineage and developmental stage of a cell type using gene expression, many aspects should be considered. In addition to considering expression of multiple genes, isoform expression adds a beneficial layer of complexity to an expression analysis. Recent reports claimed that multiple isoforms could be expressed within one cell type [181], while > 90% of human genes reportedly undergo detectable alternative splicing [182]. We found >1,200 isoform switching events within our populations,

which includes alternative splicing and differential promoter usage. Some events, such as that seen at *DNMT3B (***Figure 2**.**6***)*, suggest alternative functions for the encoded proteins given that the isoform expressed in the differentiated populations is not catalytically active [140]. We also identified expression of three *PITX2* isoforms, with differential splicing occurring between the dME and dEN. PITX2 is essential for heart looping during chick development, but the isoform expressed in the dEN does not induce the same morphologic consequence [183].

A major caveat to our analysis is that populations of cells are considered; we cannot conclude that every cell within the population expresses all transcripts detected using RNA-Seq. It is possible that multiple subpopulations exist, giving rise to one averaged transcriptional signature. Our analysis could be improved by performing single cell sequencing with considerations for small RNAs (such as enhancer RNAs or micro RNAs), non-polyadenylated transcripts and sub-cellular transcript localization transcript [184, 185]. Advances in single cell sequencing technology will provide even greater depth to our understanding of transcription in the near future. As technology improves, heterogeneity issues within population-based studies may become more apparent, forcing a new perspective on molecular regulation of cellular identity. These observations suggest that expression characteristics of combinations genes are crucial factors to consider when examining the transcriptional profile of a population.

### 5.3 Ectoderm and Endoderm Lineages Exhibit Genome-wide Similarities

Clustering analysis of the WGBS data revealed that the dME and dEN were more similar to each other than to the dEC (**Figure 3.1**). However, our transcriptional and histone modification analyses revealed that the dEC and dEN were more similar to each other than they were to the dME or HUES64, which was surprising given that the dME and dEN proceed

through a similar mesendoderm intermediate (**Figure 2.2**). When associating regions that gained H3K4me1 during differentiation with gene expression profiles of primary tissues, gain of H3K4me1 in the dEC was associated with cortex and fetal brain categories, while gain in dME was associated with heart and thyroid categories. Examining the dEN associations unexpectedly revealed that regions were associated with pancreatic islets, as well as brain-related categories such as cortex and fetal brain (**Figure 4.1**).

A GO term analysis of regions that gained H3K27ac during differentiation again returned expected gene categories for dEC (e.g. neural precursor maintenance, midbrain development and cerebellum morphogenesis) and dME (e.g. cardiac ventricle development, vasculogenesis and heart morphogenesis) (**Figure 5.1**). Remarkably, our analysis of dEN-associated GO terms returned neural tube development, dorsal spinal cord development and neuron differentiation, though the category significance did not completely overlap with the dEC, suggesting different gene sets were correlated epigenetic remodeling.

We also detected enrichment of retinoic acid signaling regulatory elements in the dEN, which is a signaling pathway associated with later stages of endoderm development. This suggests that either these regions are being remodeled to a euchromatic state in anticipation for their active role in later stages of differentiation or they are promiscuous and additionally employed in TGFβ and WNT signaling cascades. Examination of TF motifs that exhibited gain of H3K27ac during differentiation also revealed trends that correlated with the RNA-Seq and H3K4me1 analysis (**Figure 5.2**). There were few motifs that exhibited gain in both dEC and dEN, but not dME, though one notable example included the FOXA2 motif. This TF is historically associated with endoderm development and has recently been implicated in efficient

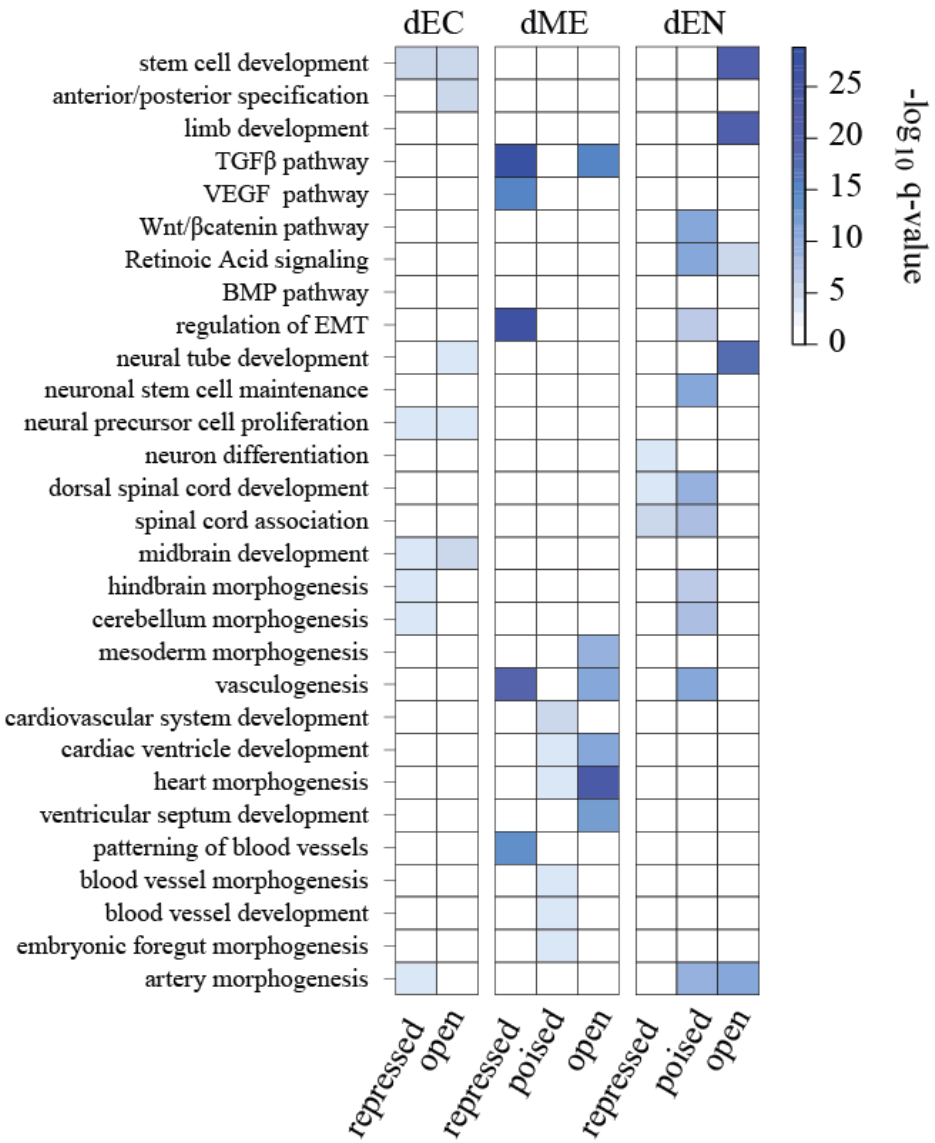



**Figure 5.1** GO categories enriched in regions transitioning to H3K27ac in the cell type indicated on the right compared to hESCs as determined by GREAT analysis. Regions gaining H3K27ac were split up by state of origin in hESC into repressed (none, IMR, HMR, and HK27me3), poised (H3K4me1/H3K27me3), and open (H3K4me3/H3K27me3, H3K4me3, and H3K4me1). Color code indicates multiple testing adjusted q value of category enrichment.

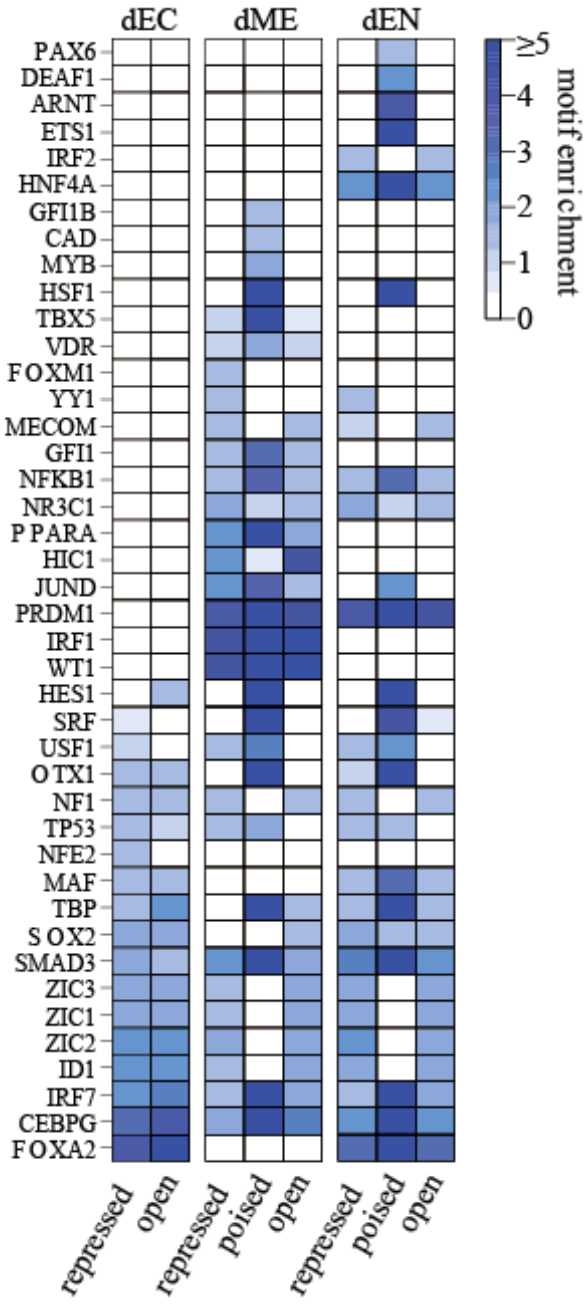



**Figure 5.2** TF motifs enriched in regions changing to H3K27ac in the cell type indicated on the right compared to hESCs. Color code indicates motif enrichment score incorporating total enrichment over background as well as differential expression of the corresponding transcription factor in the respective cell type. Regions gaining H3K27ac were split up by state of origin in hESC into repressed (none, IMR, HMR, HK27me3), poised (H3K4me1/H3K27me3), and open (H3K4me3/H3K27me3, H3K4me3, and H3K4me1).

neural differentiation of hiPSCs as well [186]. This epigenetic association was expected in the dEN given that FOXA2 is actively expressed in this population, but its significance in the dEC analysis was surprising because it is not expressed until later stages of neural differentiation. This observation again suggests that these regions may be primed such that they are prepared for activation at later stages. An alternative hypothesis is that additional factors/pathways use the same region for regulation. Our windows of histone modification enrichment were at least 400bp in length; therefore, it is possible that additional factors bind in these regions. This has some analogy to super enhancers recently identified in mESCs, where many regulators of cellular identity are bound within similar regions [187].

Altogether, the similarities between the dEC and dEN suggest that these two cell types may use similar gene regulatory elements during specification. They share the need for many transcription factors such as FOXA2, PAX6, ISL1 and NGN3, and a previous study found an overlap in epigenetic programs used in both neural tissues and *in vitro* derived β islets [188]. This may reveal novel regulatory elements that were previously associated with neural programs because an embryonic definitive endoderm population had not been extensively profiled. Our hypotheses regarding the role of H3K27ac in affecting expression level and splicing information discussed in Chapter 2 further support the idea that similar genes may be employed by each lineage, but differing between populations in expression level and the isoform expressed. Importantly, we cannot exclude the possibility that our association is due to heterogeneity in the dEN FACS sorted population or inappropriate specification of the definitive endoderm population. The underlying implication of this observation is that epigenetic profiling of *in vitro* derived populations represents a wealth of information regarding the process of human embryonic specification.

## 5.4 Transcription Factor Binding at Repressed Loci

Using binding profiles of O/S/N in HUES64, we confirmed previous reports that O/S/N binding sites maintain local depletion of DNA methylation in hESCs [10]. As expected, OCT4 binding sites frequently transitioned to a HMR state during differentiation towards all three lineages, suggesting stable silencing of the pluripotent network (**Figure 3**.**8**). Unexpectedly, we also identified many regions bound by O/S/N that are associated with genes that become expressed at later stages of development, such as *DBX1, FOXA2* and *ISL1*. Many of these regions exhibited low CpG density and gained DNA methylation during differentiation toward the lineages that do not require expression of the associated gene (**Figure 3.8**).

We also found that regions bound by O/S/N were capable of maintaining their chromatin state or transitioning to an active state during differentiation. This suggests that these TFs have multiple responsibilities in maintaining a pluripotent phenotype, including promoting an open chromatin structure at genomic regions to facilitate prompt activation at a later time, in addition to their role in positively regulating gene expression. Additional support for their association with chromatin structure maintenance was presented by a report that found these factors are commonly bound with Mediator, a protein essential for maintaining chromatin structure and DNA looping in mESCs [163].

FOXA2 binding profiles in the endoderm also suggest that this TF has functions in addition to activation of transcription. We found that many regions bound by this TF were not expressed in the dEN but they were associated with later stages of endoderm development. This suggests that FOXA2 binding alone does not sufficient to lead to immediate activation in all contexts. *In vitro* evidence suggests that FOXA proteins are capable of remodeling chromatin and influencing nucleosome positioning [176], but there are no reported interactions with

transcriptional machinery. However, it does reportedly interact with DNA-dependent protein kinase, which catalyzes phosphorylation of serine 283. This event is important for activation of FOXA2 targets [189], though phosphorylation is not required for DNA binding given that a S283A FOXA2 point mutation was capable of binding to DNA. These experiments therefore separate FOXA2's ability to bind to DNA versus activate transcription. Protein activation via phosphorylation is a common mechanism responsible for regulation of various biological processes because this post-translational modification can alter protein structure and protein-protein interactions [190].

## 5.5 Epigenetic Priming

Among many other interesting trends within our data, we observed two prevalent lineage specific transitions: high DNA methylation to either H3K4me1 or H3K27me3 enrichment (**Figures 2.9**). Our lab made similar observations for H3K4 methylation during the early stages of reprogramming to an iPSC state, suggesting that this type of epigenetic "priming" event might be common to *in vitro* cellular transitions. It was not clear from these experiments whether these events reflected a regulatory mechanism that facilitates timely activation during cellular transitions or an *in vitro* artifact that indicates the absence of a critical co-factor necessary for complete transcriptional activation *in vitro*. The latter is unlikely as profiling cardiomyocytes during murine differentiation revealed a similar trend, in which H3K4me1 is gained prior to gene expression [168].

Our observation that regions with high DNA methylation switch to H3K27me3 enrichment in CpG-poor regions bound by FOXA2 in the endoderm is one of the most interesting observations in this data set (**Figure 4.4**). It remains to be tested whether targeted loss

of DNA methylation at these regions causes a default gain of H3K27me3 in the absence of additional co-factors due to underlying sequence context or represents a more active recruitment event and regulatory mechanism [82]. It was recently reported that the combination of H3K27me3 enrichment and a nearby nucleosome-depleted region creates sites amenable to TF binding [100]. Based on these results, we speculate that specific TFs, such as FOXA2, induce chromatin decompaction that leads to loss of DNA methylation, which represents the first step in the gene activation process. The subsequent gain of H3K27me3, by default or through active recruitment, then creates a platform for subsequent binding of additional TFs that cannot directly remodel a heterochromatic state, but instead function in transcription machinery assembly and transcriptional activation [191]. Further studies should establish the extent of "priming," as this event does not lead to appreciable levels of Pol II at the AFP gene in the dEN cell type (*A. Tsankov, personal communication*), though expression is detected at the next stage of differentiation.

A comprehensive DNaseI HS mapping study also provided results supporting the basic idea that regulatory sites may be opened prior to their activation. DHS was profiled in multiple *in vivo* derived tissues, and at various stages of hESC differentiation. This study found that 56% of DNaseI HS sites found in hESCs were also detected in at least one additional differentiated population [3]. They observed gradual restriction of chromatin accessibility during differentiation, with loss being more prevalent than gain of accessibility, leading to 37% of DHS sites in terminally differentiated cell types also being identified in hESCs. The underlying implication is that a subset of the pluripotent epigenome is prepared for regulatory duty during subsequent stages of development.

The aforementioned study simultaneously highlights that many regions categorized as DHS in somatic tissues are not sensitive in hESCs, suggesting regulatory elements must acquire an accessible chromatin landscape at some time during differentiation. Recent exposure of the TET proteins suggests a mechanism that could facilitate these transitions. We have a limited understanding of the TET proteins and 5hmC because the ramifications associated with their dynamics have only recently been defined [30]. Initial efforts to study this epigenetic mechanism were complicated by the fact that traditional bisulfite sequencing does not distinguish between 5mC and 5hmC [192], and affinity purification does not provide base pair resolution [193]. A protection-based method has now been developed and found 5hmC to be localized to DHS sites and enhancer elements containing low CpG density [164].

A study of reprogramming to an iPSC state found that 5hmC appeared prior to transcriptional activation of nearby genes [194], providing the first direct evidence that 5hmC enrichment is an intermediate step during the progressive loss of DNA methylation. This result was recently confirmed in part by profiling various stages of neural development [78]. An additional report also found that 5hmC co-localized with PRC2 binding in mESCs [195], suggesting 5hmC and 5mC exhibit distinct protein interaction capabilities given that 5mC is not found at regions enriched for H3K27me3 [63]. Lastly, 5hmC did not correlate with H3K27ac in hESCs, but did show a slight preference for regions enriched for H3K4me1 [164]. Collectively, this data, in combination with our profiling, suggests that 5hmC enrichment, as well as H3K4me1 and H3K27me3 enrichment, represent epigenetic remodeling in anticipation of transcriptional activation.

An elegant study in *Drosophila melanogaster* embryos recently demonstrated the consequences of improper timing of gene expression in during embryonic development [196].

Premature or mosaic gene expression within the population prevented the rapid activation of specific transcripts necessary to promote development, which occurred when Pol II was successfully stalled at the promoter during unperturbed development. This report suggests that pausing of Pol II at promoters prior to their activation is an essential step in development, but this report made no experimental link to chromatin state. We do not detect Pol II at the *AFP* gene in the dEN cell type, suggesting this remodeling event is not sufficient for transcription initiation complex recruitment. We hypothesize that the gradual epigenetic remodeling of loci during development, may create a platform sufficient for generous Pol II accumulation that will allow swift activation in response to additional stimuli, faithfully orchestrating embryonic specification.

## 5.6 Future Directions

Our data set provides many intriguing avenues to further explore the role of epigenetic regulation during lineage specification using hESCs. Further work should be focused on establishing the necessity of priming events that occur prior to transcriptional activation. Ectopic reporter constructs similar to those used to establish sequence requirements for DNA methylation [26] could be constructed to similarly mimic and perturb chromatin remodeling, to separate sequence context driven events from indirect events due to looping [197, 198]. Studying loci in isolation, ectopically, would allow us to investigate FOXA2 directed remodeling, at the AFP gene for example, without limiting the differentiation potential of the population. If the reporter shows decreasing methylation at FOXA2 binding sites, that suggests binding and loss of methylation is sequence-dependent. Mutations in the binding sites could then establish the sequence requirement for binding and demethylation, and subsequent gain of H3K27me3. Another avenue of pursuit would be the use of TAL effectors that are attached to histone

modifying enzymes, which can then be targeted to specific regions throughout the genome [199]. The combination of these two approaches will provide novel insight in to the epigenetic remodeling that is necessary for proper gene regulation during differentiation.

We are interested in separating 3-dimensional architecture from direct binding given that multiple FOXA2 peaks are often clustered close together. This curiosity is fueled by our observation that the CpG methylation level at a subset of peaks does not decrease during differentiation, and this frequently occurs when multiple peaks are detected within close proximity to each other (unpublished observation). It is possible that the CpGs exhibiting demethylation represent the true FOXA2 binding sites, while the nearby peaks represent secondary peaks due to looping interactions. The alternative interpretation is that FOXA2 can bind regions containing methylated DNA. These two conclusions are indistinguishable in the current data set. It is widely believe that the ability to bind 5mC is an uncommon feature of TFs; therefore, understanding which TFs maintain this affinity is essential. Discovering which TFs can bind methylated DNA will likely reveal regulators that induce cellular transitions. In parallel, high throughput reporter assays may additionally confirm the regulatory potential of enhancers that are bound by these TFs and experience loss of DNA methylation [200].

Inference regarding direct versus indirect binding can also be made using DNaseI HS data in combination with ChIP-Seq [201]. By combining TF-ChIP peaks, TF motifs and DHS sites, a recent study examined the correlation between these factors. Regions that lacked a DHS site or motif consistently exhibited lower enrichment based on ChIP-Seq, suggesting they were indirect sites of binding, likely crosslinked together because of close proximity in space rather than as a result of direct binding. DHS profiling in our cell types would therefore improve our ability to understand direct protein-DNA interactions, but also reveal general sites of protein-

DNA interactions that are lineage specific. The loss of DNA methylation at many loci, including *POU3F1* in the dEC, was not well covered in the ENCODE data, suggesting that many regulatory elements associated with development may currently be undisclosed.

Much attention has been directed towards chromosome architecture with the development of assays such as DNA adenine methyltransferase identification (DamID) and chromosome conformation capture techniques that promote an understanding of nuclear localization and 3-dimensional genomic architecture. In DamID, the prokaryotic adenine methyltransferase *dam* was fused to Lamin B1, which is located in the nuclear envelope, which results in the appearance of adenine-6-methylation when the fusion protein interacts with DNA [202]. This approach gave way to the idea that repressed chromatin domains, such as those enriched for H3K27me3 and H3K9me2, were located in the nuclear periphery [202]. Hi-C was later designed to use proximity ligation as a means to study three-dimensional genomic architecture [203]. This approach captures interactions throughout the genome, but suffers in resolution, making it difficult to confidently make conclusions about discrete genomic interactions. 4-C takes a step back from Hi-C, and interrogates the chromosomal interactions of one locus [197]. These approaches would be useful for understanding the architecture associated with the HMR-to-H3K27me3 switch, as well as general nuclear reorganization that accompanies specification.

The mechanism by which FOXA2 induces decompaction could also be rectified by studies of the Tet proteins and hydroxymethylcytosine enrichment. TET1 was recently labeled as a Nanog interacting partner during reprogramming, and the Tet-related catalytic activity was shown to facilitate epigenetic priming at pluripotent loci [194]. This report also suggested the 5hmC enrichment was detected prior to transcriptional activation, therefore defining an additional epigenetic modification that discerns primed loci. During differentiation, it is therefore

possible that FOXA2 binding recruits TET proteins to HMRs that initiates loss of DNA methylation. FOXA2-directed decompaction may allow TET binding based on chromatin accessibility, or FOXA2 may directly interact with and recruit the TET proteins. An additional prospective mechanism is that TET-directed oxidation to 5hmC occurs first, subsequently allowing FOXA2 binding. Until discovery of the role for Tet proteins in DNA demethylation, few studies had convincingly addressed the process by which regions containing high DNA methylation transition to active states during differentiation and development. 5hmC profiling in our system would likely provide numerous exciting observations regarding these events.

Given our observation that splicing causes expression of a catalytically inactive DNMT3B protein (DNMT3B3), we have renewed interest in the function of this protein during differentiation and development. The original study that reported embryonic lethality during embryonic development interrupted the catalytic domains, creating a catalytically inactive allele. Heterozygous mice appeared to develop normally, though homozygous mice exhibited slowed growth and neural defects at E11.5. Crossing of two heterozygous *Dnmt3b* null mice did not give rise to live homozygous pups in this study [20]. The inactive isoform is reportedly expressed in various transformed cell lines and at low levels in some somatic tissues [138, 204], as it is quickly downregulated during post-implantation development [20].

*Dnmt3b* deletion in mESCs does not inhibit self-renewal and proliferation and does not cause loss of DNA methylation at satellite regions during prolonged *in vitro* culture. Alternatively, deletion of both *Dnmt3a* and *Dnmt3b* in mESCs leads to global loss of DNA methylation, which can be restored by reintroduction of the active *Dnmt3b* gene, but not the catalytically inactive gene that is expressed in each of our differentiated populations. These observations suggest that DNMT3B3 may exert functions other than de novo DNA methylation

during embryonic development or that DNMT3B de novo activity is not required after gastrulation. The lethality observed in Okano et al may be the result of impaired activity that occurs pre-implantation when the active isoform is expressed, but the phenotype may take many more cell divisions to manifest, resulting in lethality at a later stage.

To understand regions subject to aberrant regulation due to defects in DNMT3B, we are currently examining recently published WGBS data from lymphoblasts of an ICF syndrome patient that exhibits a *DNMT3B* mutation [205]. This study found depletion of DNA methylation at centromeric satellites as expected and 42% overall reduction of DNA methylation on autosomes. Interestingly, regions associated with H3K36me3 and H3K79me2 over actively transcribed gene bodies remained highly methylated. The largest decrease in DNA methylation was seen at CpG poor promoters and intergenic regions associated with H3K4me1 in the WT patient. By further examining this data, we can begin to dissect regions that require DNMT3B for proper regulation and try to identify the developmental stage at which they require DNMT3B. Going forward, using ICF patient cell lines that exhibit deficient DNMT3B activity offers an excellent model system for understanding the interplay between DNA methylation and histone modifications.

While we cannot confirm that the remodeling observed in our system occurs *in vivo*, understanding the epigenetic remodeling that facilitates specification can likely improve efforts to produce functional tissues *in vitro*. Many current *in vitro* differentiation protocols produce cell types that appear to be in a premature state [132, 206]. Our observations suggest that a cell type exhibits epigenetic remodeling that forecasts the transcriptional activation destined to occur at later stages (e.g. *AFP*). Similarly, understanding the chromatin state at multiple regulatory elements (e.g. *NODAL*) required for gene activation is important for understanding a

population's potential. Therefore, comparing the epigenetic landscape of immature cells produced *in vitro* to primary tissues, may provide regulatory insight that will aid in the production of therapeutically relevant cell types *in vitro*.

# Chapter 6.

# Methods

6.1 Cell Culture

All *in vitro* derived cell types were derived from HUES64 [129]. Human embryonic stem cells were expanded on murine embryonic fibroblasts (Global Stem) in KO-DMEM (Life Technologies) containing 20% Knockout serum replace (Life Technologies) and FGF2 (10 ng/mL) (Millipore). Cultures were passaged by enzymatic dissociation using Collagenase IV (1mg/mL) (Life Technologies). Prior to differentiation, cells were plated on matrigel-coated plates (BD Biosciences) and cultured in mTeSR1 (Stem Cell Technologies) for 3 to 4 days. Endoderm differentiation was induced in Advanced RPMI (Invitrogen), 0.5% FBS (Hyclone), Activin A (100ng/mL) (R&D) and WNT3A (50 ng/mL) (R&D). HUES64- derived hepatoblasts (dHep) were induced by culturing day 5 endoderm in RPMI media containing B27(1X), FGF2 (10ng/mL)(Millipore) and BMP4(20ng/mL)(R&D) for five days, and collected after 10 days total of differentiation. Hepatocyte-like cells were derived by culturing the HUES64-derived hepatoblasts in Lonza hepatocyte culture media containing 10ng/mL of HGF (R&D) for 5 additional days, or 15 days total. Mesoderm differentiation was induced by the addition of media consisting of in DMEM/F12 (Life Technologies), 0.5% FBS (Hyclone), Activin A (100ng/mL) (R&D) (for the first 24 hours only), BMP4 (100ng/mL) (R&D), VEGF (100ng/mL) (R&D) and FGF2 (20ng/mL) (Millipore). To induce osteoblast differentiation, the day 5 mesoderm population was dissociated with accutase and replacted on matrigel coated plates (BD) in EGM-2 media (Lonza) for 7 days, or 12 days total. Ectoderm differentiation was induced using A83-01 (2um) (Tocris), PNU 74654 (2um) (Tocris) and Dorsomorphin (2um) (Tocris), DMEM/F12 (Life Technologies) containing 15% Knock serum replacer (Life Technologies). Neurectoderm differentiation was induced by switching the day 5 ectoderm population to media containing 3 μM CHIR99021 (TOCRIS), 10 μM SU5402 (TOCRIS), and 10 μM DAPT (TOCRIS), and

collected after 6 more days, or 11 days total. N2-supplement (Life Technologies) was added to cells in 25% increments every other day beginning four days after the initiation of ectoderm differentiation. For all cell types, media was changed daily.

## 6.2 NanoString Analysis

The same probe set previously designed by our lab to interrogate regulators of development was employed here [129]. RNA was isolated using the Qiagen RNeasy kit. 500ng of whole RNA was incubated with the probe set for 18 hours at 65°C. The analysis was done as previously described.

## 6.3 Antibodies

ChIP was performed using the following antibodies: H3K4me3 (Millipore, 07-473, Lot DAM1623866), H3K27ac (Abcam, ab4729, Lot 509313), H3K27me3 (Millipore, 07-449, Lot DAM1514011), H3K36me3 (Abcam, ab9050, Lot 499302), H3K4me1 (Abcam, ab8895, Lot 659352), H3K9me3 (Abcam, ab8898, Lot 484088), POU5F1 (Abcam, ab19857), SOX2 (Santa Cruz, sc-17320X), NANOG (R&D, AF1997) and FOXA2 (R&D, AF2400).

For live cell FACS isolation, cells were stained for 30 minutes on ice with the following antibodies directed towards extracellular surface proteins: CD326-PerCP-Cy5 (clone EBA1) (BD Biosciences), CD56-PE (clone NCAM16.2) (BD Biosciences), and CD184-PE-Cy5 (clone 12G5) (BD Biosciences).

Immunostaining was done with the following primary antibodies: FOXA2 (R&D, AF2400), GATA2 (Santa Cruz, sc-16044) SOX17 (R&D, AF1924), PAX6 (Covance, PRB-278P) and HNF4α (abcam, ab41989). Cells were fixed in 4% Formaldehyde, incubated in

primary antibody overnight at 4°C, and then incubated in secondary antibody for 1 hr at room temperature. DNA was detected using Hoechst 33342 trihydrochloride trihydrate (Invitrogen).

## 6.4 FACS Analysis

FACS was done on a BD FACSAria™ II using linear FSC and SSC scaling, followed by height and width-based doublet discrimination. The viability of the populations was assessed by Propidium Iodide staining, with the positively stained populations being excluded from the sorting gates. Compensation was calculated using FACS Diva autocompensation algorithms, and supplemented by manual compensation to correct for autofluorescence. Antibodies were used as described in the main text.

## 6.5 WGBS-related Protocols

### Genomic DNA isolation

Flash-frozen human tissues or cell pellets were lysed at 55°C overnight in 300-600 µl lysis buffer (10 mM Tris-HCl pH 8.0, 10 mM EDTA, 10 mM NaCl and 0.5% wt/vol SDS) supplemented with 50 ng/µl DNase-free RNase (Roche) and 1 µg/µl proteinase K (NEB). After extraction with an equal volume of phenol:chlorofom:isopropanol alcohol (25:24:1; Invitrogen) and addition of 0.5 µl (20 µg/µl) glycogen (Roche) and 1/20 vol 5 M NaCl, DNA was precipitated with 2.5 vol ethanol, spun down (30 min/16,000 g) at 4°C and washed with 70% ethanol. DNA was re-suspended in 30-100 µl of TE buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA) and quantified using a Qubit fluorometer and a dsDNA BR Assay Kit (Life Technologies).

WGBS Library Construction

Genomic DNA (1-5 µg) was fragmented to 100-500 bp using a Covaris S2 sonicator 9 times for 60 s at duty cycle 20%, intensity 5 and 200 cycles per burst. DNA fragments were cleaned up using a QIAGEN PCR purification kit. End-repair reactions (100 µl) contained 1x T4 DNA ligase buffer (NEB), ATP, 0.4 mM dNTPs, 15 units T4 DNA polymerase, 5 unites Klenow DNA polymerase, 50 units T4 polynucleotide kinase (all NEB) and were incubated for 30 min. at 19°C and 15 min. For some libraries we used a dCTP-free dNTP mix instead of all four dNTPs during for the end-repair to avoid artificially unmethylated sites. Adenylation was performed for 30 min. at 37°C in 50 µl 1x Klenow buffer containing 0.2 mM dATP and 15 units Klenow exo$^-$ (NEB). Adenylated DNA fragments and methylated paired-end adapters (purchased from ATDBio) were incubated overnight at 16°C in a 50 µl reaction containing 5,000 units concentrated T4 DNA ligase (NEB) and 3 µM of adapters. Each enzymatic reaction was terminated and cleaned-up by phenol/chloroform extraction and ethanol precipitation as described above.

To determine unmethylated cytosine conversion rates and methylated cytosine over-conversion rates by sodium bisulfite treatment, adapter-ligated fully methylated and fully unmethylated internal control DNA fragments, were spiked into WGBS library preparation at a molar ratio (spike-in to WGBS library) of 1:16,000 each. Adapter-ligated DNA of 270-370 bp, corresponding to DNA insert sizes of 150-250 bp, was size-selected on a 2.5% Nusieve (3:1) agarose gel (Lonza). Two consecutive bisulfite conversions were performed with an EpiTect Bisulfite Kit (QIAGEN) following the protocol specified for DNA isolated from FFPE tissue samples. One of 40 µl bisulfite-converted DNA was used in each of four 10-µl reactions to determine the minimal PCR cycle number for library amplification. PCR reactions contained 0.5

U of PfuTurboCx Hotstart DNA polymerase (Agilent technologies), 1 µl of 10x PCR buffer, 250 µM dNTPs, 1.5 µM of Primer 1.0 and 2.0 (Illumina). The thermocycling profile was 2 min. at 95°C followed by 5-15 cycles of 30 s at 95°C, 30 s at 65°C, 1 min. at 72°C, and a final 7-min. extension at 72°C. Preparative library amplification using the empirically determined number of PCR cycles was performed in eight 25-µl aliquots, each containing 3 µl of bisulfite-converted DNA, 1.25 U of PfuTurboCx Hotstart DNA polymerase, 2.5 µl of 10x PCR buffer, 250 µM of dNTP, 1.5 µM of Primer 1.0 and 2.0. PCR products were pooled and purified twice using Agencourt AMPure XP SPRI Beads (Beckman Coulter) as per the manufacturer's instructions. The final library DNA was quantified using a Qubit fluorometer and a Quant-iT dsDNA HS Kit (Invitrogen). The insert size was checked on a 4-20% non-denaturing polyacrylamide gel (Bio-Rad). Paired-end sequencing with 100 base reads was performed on an Illumina Hiseq 2000 followed the manufacturer's guidelines.

WGBS Data Processing and Analysis

WGBS raw sequencing reads were aligned using maq in bisulfite mode against human genome version hg19/GRCh37, discarding duplicate reads. DNA methylation calling was performed based on an extended custom software pipeline published previously for RRBS [129]. To ensure comparability of region DNA methylation levels across all samples, only CpGs covered by ≥5x in 85% of the samples qualified for the computation of region DNA methylation levels. To assess the DNA methylation state of various genomic regions, we resorted to our previously published protocol estimating a genomic region's methylation state as the coverage weighted average across all CpGs within each region. Subsequently, we averaged a region's DNA methylation level over replicates. Differentially methylated regions (DMRs) were defined

as exhibiting significantly ($p \leq 0.05$, fisher's exact test) different DNA methylation levels of at least 0.1.

Many gene regulatory elements (GREs) are marked by spatially highly constrained reduced DNA methylation levels. It has recently been suggested that besides CpG islands, which are mostly unmethylated (UMR) a second class of GRE is marked by low to intermediate DNA methylation (IMR). We reasoned that these regions might be of particular regulatory importance in our system and might be missed by looking at histone modification enrichments alone. Therefore we adopted a similar Hidden Markov model approach as proposed in Stadler et al. to identify regions of reduced DNA methylation level. Briefly, we utilized a three state Hidden Markov Model operating on the methylation levels of each CpG in the human genome. Each state's emission probabilities for the DNA methylation levels were modeled by a normal distribution. The model was trained on all CpGs of chromosome 19 in the HUES64 dataset using an adaption of the well known Baum Welch algorithm to incorporate the normal distribution [207]. After initial parameter estimation, we utilized the approach reported by Stadler et al. to determine the FDR for IMR regions and adapted the initial parameter estimates for the IMR and HMR states to finally 0.01(UMR), 28.8 (IMR), 81.6 (highly methylated, HMR), yielding an FDR of 2%. This parameter set was subsequently used to segment all WGBS datasets. Finally, we used the Viterbi algorithm to compute the most probable path through each chromosome separately and assigned the CpG states accordingly to either unmethylated, intermediate or highly methylated. Subsequently, we merged neighboring CpGs residing in the same state and being less than 200bp apart into unmethylated, intermediate or highly methylated regions. Only regions harboring more than 3 CpGs were retained for subsequent analysis. The resulting region set is more likely to pick up DMRs due to the highly spatially constrained nature of the marked

GRE (often 200-400bp) which easily gets masked by a coarse grained tiling based approach. The HMM inference framework was implemented as custom software in python (http://python.org/) and extended to incorporate other state distribution types. To determine differentially methylated regions between two samples, we followed our previously established protocol [129].

## 6.5 ChIP-related Protocols

### ChIP and ChIP-Sequencing Library Production

Cells collected by FACS were crosslinked in 1% formaldehyde for 10 minutes at room temperature, with constant agitation, followed by quenching with 125mM Glycine for 5 minutes at room temperature with constant agitation. Nuclei were isolated and chromatin was sheared using Branson sonifier until the majority of DNA was in the range of 200-700 base pairs. Chromatin was incubated with antibody overnight at 4°C, with constant agitation.

Co-immunoprecipitation of antibody-protein complexes was completed using Protein A or Protein G Dynabeads for 1 hour 4°C, with constant agitation. ChIPs were completed using previously reported methods [144]. Sequencing library production details can be found in the Supplemental Experimental Procedures. Sequencing libraries were submitted for sequencing on the Illumina Hiseq 2000.

Immunoprecipitated DNA was end repaired using the End-It DNA End-Repair Kit (Epicentre), extended using a Klenow fragment (3'-5' exo)(NEB), and ligated to sequencing adapter oligos (Illumina). Each library was then PCR-amplified using PFU Ultra II Hotstart Master Mix (Agilent), and a size range of 300-600 was selected for sequencing.

ChIP-Bisulfite Sequencing Library Construction

DNA was first subjected to end-repair in a 30-µl reaction containing 6 units T4 DNA polymerase, 2.5 units DNA Polymerase I (Large Klenow Fragment), 20 units T4 Polynucleotide Kinase (all New England Biolabs), dATP, dCTP, dGTP, and dTTP (0.125 mM each), and 1× T4 Ligase buffer with ATP for 30 min at 20°C. DNA was then adenylated in a 20-µl reaction containing 10 units Klenow Fragment (3′→5′ exo-) (New England Biolabs), 0.5 mM dATP and 1× NEB buffer 2 for 30 min at 37°C. DNA was then ligated to preannealed Illumina genomic DNA adapters containing 5-methylcytosine instead of cytosine (ATDBio) using T4 DNA ligase (New England Biolabs).

Adapter-ligated DNA fragments were subsequently purified by phenol extraction and ethanol precipitation and size-selected on gel. 50 ng sheared and dephosphorylated Escherichia coli K12 genomic DNA was added to adapter-ligated DNA as carrier during size-selection and bisulfite conversion. DNA was run on 2.5% Nusieve 3:1 Agarose (Lonza) gels. Lanes containing marker (50 bp ladder; New England Biolabs) were stained with SYBR Green (Invitrogen), and size regions to be excised were marked with toothpicks and adapter-ligated DNA fragments from 200–400 and 400–550 bp were excised. DNA was isolated from gel using the MinElute Gel Extraction kit (QIAGEN). The low and high libraries were kept separate in subsequent steps.

Adapter-ligated and size-selected DNA was subjected to two subsequent 5-h bisulfite treatments using the EpiTect Bisulfite kit (QIAGEN) following the manufacturer's protocol for DNA isolated from FFPE tissue samples. PCR amplification was done with 1.25 units Pfu Turbo Cx Hotstart DNA Polymerase (Stratagene), primer LPX 1.1 and 2.1 (0.3 µM each), dNTPs (0.25 mM each), 1× Turbo Cx buffer. Amplified libraries were purified with the MinElute PCR

Purification kit (QIAGEN) and subsequently purified from gel essentially as described above; whole gels were stained with SYBR Green, and no carrier DNA was added. Final libraries were analyzed on analytical 4%–20% TBE Criterion precast gels (BioRad), and measured by Quant-iT dsDNA HS Assays (Invitrogen).

ChIP-Seq Data Processing and Analysis

ChIP-Seq data was aligned to the hg19/GRCh37 reference genome using bwa version 0.5.7 [208] with default parameter settings. Subsequently, reads were filtered for duplicates and extended by 200bp. Visualization of read count data was performed by converting raw bam files to .tdf files using IGV tools [209] and normalizing to 1 million reads.

In order to identify regions enriched for chromatin modifications we employed a two step approach, first identifying all regions enriched for any chromatin modification. Next, using this comparatively small region set, we determined the quantitative enrichment level as well as significance of enrichment using a Poisson background model based on the whole cell extract (WCE). Finally, we utilize conservative enrichment and significance cutoffs to binarize our enrichment signal in order to increase robustness and simplify subsequent analysis.

First, we segmented the genome into non-overlapping windows and classified each window into either enriched or not enriched. This analysis was conducted separately for two groups, 1; H3K27ac, H3K4me3, using 200bp windows and 2; H3K27me3, H3K9me3, K3K4me1 using 400bp windows. To compute the enrichment statistics on the window level, we determined the number of unique insert size extended sequence tags whose midpoint was located within the window of interest for the ChIP-Seq track of interest as well as the WCE. Next, we used the Poisson model proposed in Mikkelsen et al 2012, to determine nominal p-value of enrichment

and computed the enrichment over the WCE. Only windows enriched at a significance level below $p< 10^{-5}$ and an enrichment above background greater than 3 was retained. For most enrichment analysis we employed only the replicate with the strongest signal.

Next, enriched windows within a distance of 850bp were merged into larger regions. Regions smaller than 400bp (600bp for broad marks) after merging were discarded as due to noise and regions greater than 10kb were split. This procedure was carried out for three groups of histone ChIP-Seq tracks separately: H3K4me3 &H3K27ac, H3K4me1 and H3K27me3 & H3K9me3 across all 4 cell types. The resulting three lists of enriched regions were then merged in a hierarchical fashion: first regions identified based on H3K4me3&H3K27ac and H3K4me1, retaining all H3K4me3 & H3K27ac regions but merging or splitting enriched H3K4me1 regions.

After completion of this initial processing step, regions were again filtered for minimal size discarding regions smaller than 400bp. Next, the same procedure was repeated for the new H3K4me3, H3K27ac, H3K4me1 region set and the H3K27me3, H3K9me3 region list. Finally, the resulting list was merged with the regions classified as UMRs and IMRs, adding only regions not overlapping with any region identified so far. This procedure gave rise to the region catalog used in subsequent analysis.

In the second processing step, comparative analysis of ChIP-Seq experiments and assignment of chromatin states was carried out, excluding regions enriched for H3K9me3 only. First, for each region in the region catalog the significance and enrichment over WCE was determined using Poisson statistics applied to the duplicate filtered and insert size extended sequencing tag counts overlapping each identified region. Regions with tag counts deviating at a significance level of $p<0.001$ from the WCE and exhibiting enrichment over WCE $\geq 3$ were classified as enriched. We chose these moderately stringent thresholds in order also pick up

chromatin state changes that occur only in a subset of the investigated cell population and therefore have lower signal. However, this comes at the expense of a higher false positive rate. Next, we compared the enrichment levels for all four cell types (hESC, dEC, dME, dEN) for each epitope separately. To that end we used the Poisson model based approach proposed in and defined regions deviating by ≥3 fold at a significance level of $p \leq 0.05$ as being different. Next, we reconciled these differential enrichment calls with our enrichment over background classification. Since in our setting we were mostly concerned with incorrectly called differences between cell states (false positives) due to heterogeneity in the distinct populations and varying ChIP-Seq library complexity, we redefined regions that were classified as enriched in hESC and not enriched in one of the differentiated cell types but exhibiting no significant difference according to our differential analysis as being enriched in the differentiated cell type under study. This approach yields a lower false positive rate in terms of dynamics at the expense of a higher false negative rate. However, at this point it still remains to be determined what magnitudes of differences in chromatin modifications are actually meaningful. In this sense, our binary classification approach is rather conservative and relies on previously established observations. Subsequently, we classified each genomic region identified in this way into one of 11 epigenetic states based on the binary classification of enrichment levels for the various modifications. DNA methylation levels were not taken into account when histone modification based states were assigned. Only states devoid of significant enrichment for one of the histone modifications were classified based on DNA methylation levels. Genomic regions were associated with their nearest RefSeq gene using the R package ChIPpeakAnno [210] and classified into promoter, intragenic, distal (<50kb from TSS and not promoter) and intergenic.

TF ChIP-Seq Analysis

For OCT4, SOX2, NANOG and FOXA2 aligned read files were processed with macs version 1.4 [211] using the following parameters: -g 2.7e9 --tsize=36 --pvalue=1e-5 --keep-dup=1 and the HUES64 WCE as input control. All other parameters were left at their default setting. For our 25bp libraries, tsize was set to 25. FDR was calculated using macs built in function essentially comparing the original read count distribution with a randomly shuffled distribution. Following this initial peak calling, only peaks significant at an FDR of 0.05 and present in both replicates were retained. As a second replicate for our OCT4 ChIP-Seq experiment we took advantage of publically available OCT4 data [212].

ChIP Bisulfite Sequencing Analysis

For the FOXA2 ChIP-bisulfite sequencing experiment, the bisulfite treated ChIP library was processed similarly to the WGBS processing described above and subsequently overlaid with the peak calling results from the FOXA2-ChIP-Seq library that was not bisulfite treated.

Motif Analysis

Predefined sets of genomic regions were scanned for occurrences of motifs contained in the Transfac professional database (2009) using the FIMO program from the MEME suite [213]. Only motifs with at least one known associated human transcription factor and detected at a significance level of $p \leq 10^{-5}$ were used for further analysis. Next, the total number of occurrences was calculated for each motif. To correct for sequence composition, we trained a Hidden Markov Model on each set of input sequence sets and generated 10 sets of number and size matched region sets using the inferred probabilities as controls. Subsequently, these sequence sets were

also subjected to the same motif identification procedure and motif enrichment results were averaged over the 10 control runs. We defined the final motif enrichment score as the fraction of total motif occurrences in the region set of interest and the total number of motif occurrences in the averaged control region set. To determine differentially enriched motifs between region sets from different hESC-derived cell types, we calculated the fraction of motif scores between the two conditions, retaining only motifs with a differential enrichment ≥1.2.

For the H3K27ac motif analysis, we computed overall motif enrichment scores for each region class separately as described above. Next, we correlated the motif enrichment scores only focusing on those motifs with scores ≥1.2. To that end we multiplied the motif enrichment score for the cell type of interest with the $\log_2$ fold change of the associated transcription factor in that cell type, giving rise to a new combined motif score. If multiple TFs mapped to one motif, we took the average motif score.

For the H3K4me1 analysis, we wanted to focus on all potential TFBS gaining H3K4me1 and not only those that also become expressed as in the H3K27ac analysis. First, we again determined the motif enrichment scores over background. To focus on motifs differentially enriched between the different cell types, we subtracted the mean motif enrichment across the differentiated cell types for each motif separately from the enrichment level and rank ordered the motifs.

For the analysis of potential upstream regulators of transcription factors that exhibit changes in their promoter region, we first scanned all distal and proximal regions associated with these transcription factor genes for motif occurrences. Next, we determined whether any of the observed motifs were associated with transcription factors differentially expressed in any of the cell types and correlated the sign of the differential expression in each cell type with the sign of

the epigenetic state change of the region the motif occurred in (gain of open chromatin mark: +1, loss of open chromatin mark or acquisition of a repressive state: -1). Next, we rank ordered all observed TF motifs that were differentially expressed in at least one of the cell types based on their occurrence/epigenetic state change correlation for each cell type separately and reported the gene expression levels of top 30 motifs for each cell type.

## 6.6 RNA-Seq Related Protocols

### RNA-Seq Library Generation

Polyadenylated RNA was isolated using Oligo dT beads (Invitrogen). The Poly-A fraction was then fragmented to 200–600 base pairs and then ligated to RNA adaptors using T4 RNA Ligase (NEB), preserving strand of origin information. cDNA was then reverse transcribed using a universal primer that annealed to the RNA adaptor. Then RNA was degraded, and the cDNA was ligated to a DNA adaptor. Final library amplification was done using bar-coded primers that annealed to this DNA adaptor.

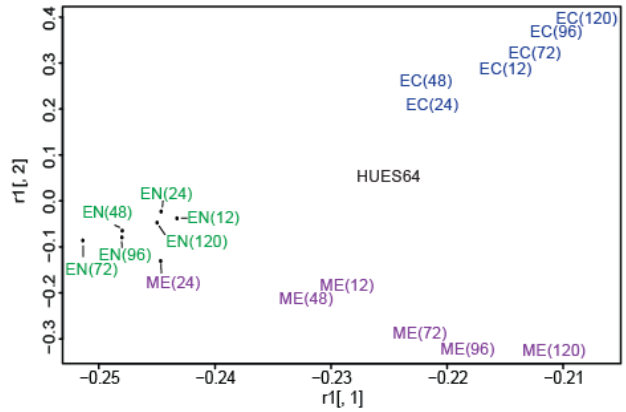
### RNA-Seq Data Processing and Differential Expression analysis

Strand specific libraries were constructed as described in the main text using a strand specific method [214]. Reads were mapped to the human genome (hg19) using TopHat v2.0.6 [215] (http://tophat.cbcb.umd.edu) with the following options: "—library-type firststrand" and "—transcriptome-index" with a TopHat transcript index built from RefSeq. Transcript expression was estimated with an improved version of Cuffdiff 2 [135] (http://cufflinks.cbcb.umd.edu). Cuffdiff was run with the following options: "—min-reps-for-js-test 2 –dispersion-method per-condition" against the UCSC iGenomes GTF file from Illumina

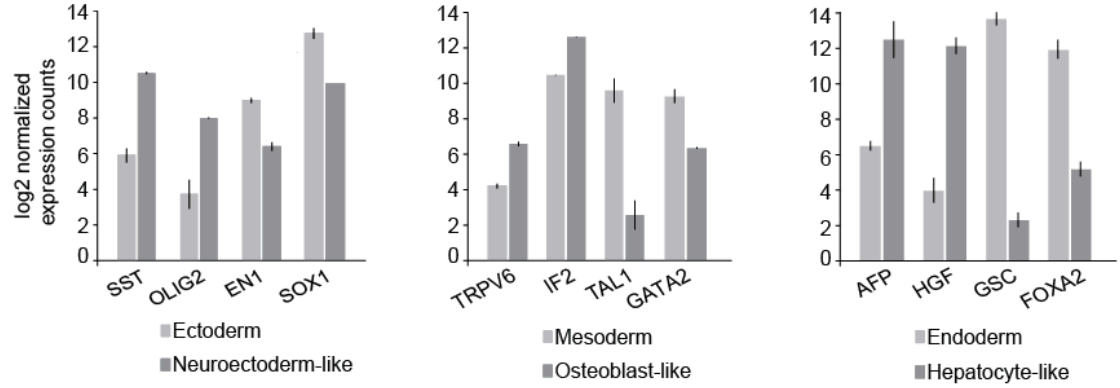
(available at http://cufflinks.cbcb.umd.edu/igenomes.html). The workflow used to analyze the data is described in detail in [216] (alternate protocol B).

# Appendix

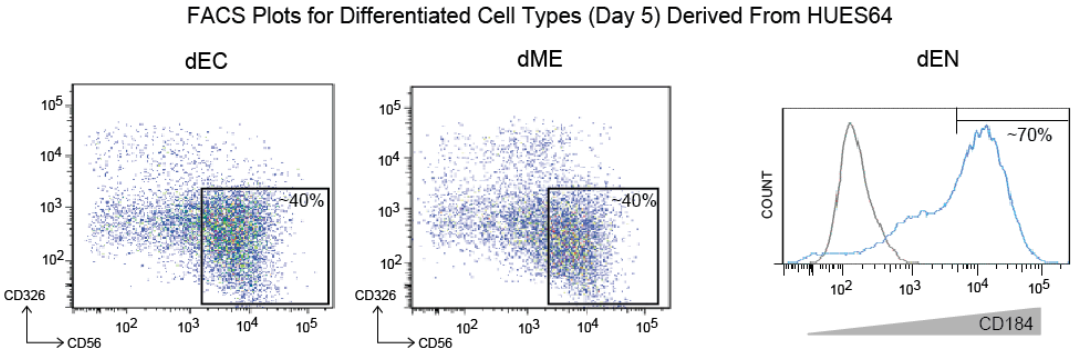
## Supplemental Figures



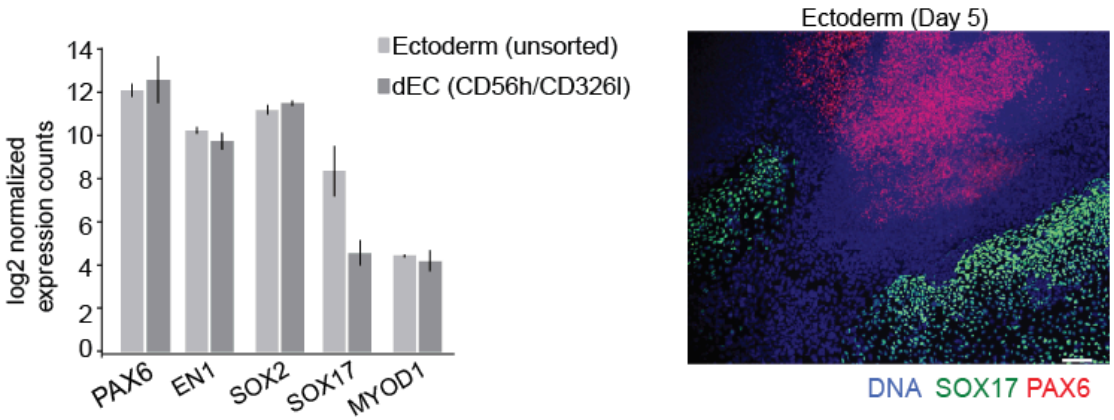

S1. Multidimensional scaling of populations included in the differentiation time course.



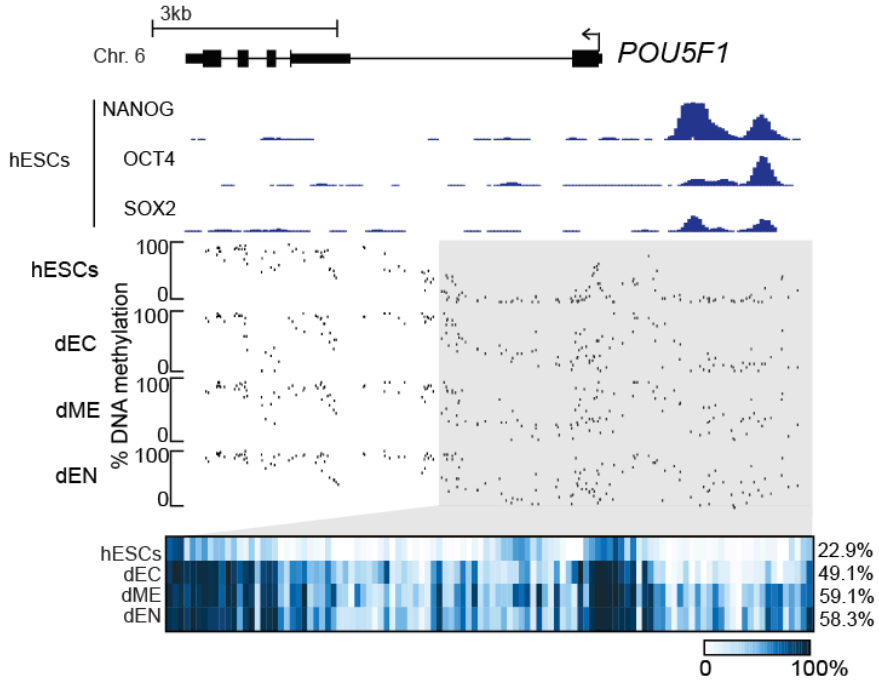

S2. Characterization of the Differentiated Populations. Median Nanostring expression values ($log_2$) of populations derived from dEC, dME and dEN.



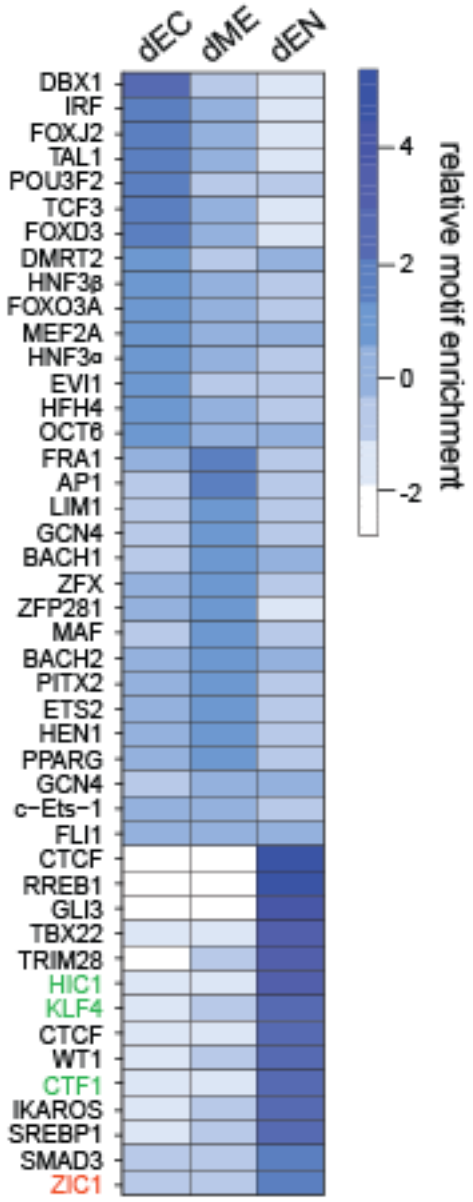

S3. Representative FACS plots used to isolate differentiated populations. Square boxes (left and middle panels) and line (in right panel) indicate population collected for further analysis. Approximate percent of population collected is given.



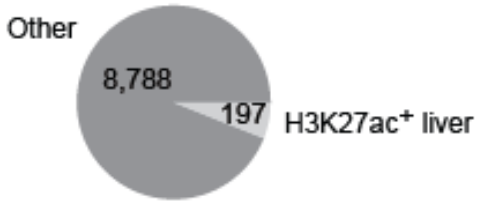

S4. Average Nanostring expression values ($log_2$) of unsorted ectoderm versus CD56high/CD326low sorted dEC cells. Immunofluorescent staining of SOX17 and PAX6 in day 5 ectoderm population (40x, scale bar equals 200um).




S5. DNA Methylation Dynamics at *POU5F1* and Regions Associated with TF Binding, Related to Figure 3(A) Gain of DNA methylation at the *POU5F1* locus (chr6:31,135,410-31,141,237). NANOG, OCT4, SOX2 ChIP-seq tracks (hESCs only) and DNA methylation levels in hESCs and differentiated cell types. Individual CpG methylation values across the locus are displayed using the IGV. The heatmap below shows the DNA methylation values of individual CpGs within the gray region. The average DNA methylation value for the entire highlighted region is shown on the right in red. The TSS is indicated by the arrow. Gain of DNA methylation is seen at the distal enhancer, as well as over the TSS, in all three differentiated cell types.




S6. Normalized motif enrichment scores for the top 15 motifs enriched in regions specifically transitioning to H3K4me1 in the differentiated cell type indicated on the bottom. Motif highlighted in green corresponds to a TF that is upregulated at the next stage (hepatoblast) of endoderm differentiation, whereas motifs highlighted in red are specifically upregulated in dEN but are downregulated at the dHep stage.




S7. Fraction of regions gaining H3K27me3 in dEN and being enriched for H3K27ac in human liver (n = 197).

# References

1. Gifford, C.A. and A. Meissner, *Epigenetic obstacles encountered by transcription factors: reprogramming against all odds.* Curr Opin Genet Dev, 2012. **22**(5): p. 409-15.

2. Hemberger, M., W. Dean, and W. Reik, *Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal.* Nat Rev Mol Cell Biol, 2009. **10**(8): p. 526-37.

3. Stergachis, A.B., et al., *Developmental fate and cellular maturity encoded in human regulatory DNA landscapes.* Cell, 2013. **154**(4): p. 888-903.

4. Gaspar-Maia, A., et al., *Open chromatin in pluripotency and reprogramming.* Nat Rev Mol Cell Biol, 2011. **12**(1): p. 36-47.

5. Arnold, S.J. and E.J. Robertson, *Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo.* Nat Rev Mol Cell Biol, 2009. **10**(2): p. 91-103.

6. Cantone, I. and A.G. Fisher, *Epigenetic programming and reprogramming during development.* Nat Struct Mol Biol, 2013. **20**(3): p. 282-9.

7. Thomson, J.A., et al., *Embryonic stem cell lines derived from human blastocysts.* Science, 1998. **282**(5391): p. 1145-7.

8. Bird, A., *The dinucleotide CG as a genomic signalling module.* J Mol Biol, 2011. **409**(1): p. 47-53.

9. Ehrlich, M., et al., *Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells.* Nucleic Acids Res, 1982. **10**(8): p. 2709-21.

10. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.

11. Stadler, M.B., et al., *DNA-binding factors shape the mouse methylome at distal regulatory regions.* Nature, 2011. **480**(7378): p. 490-5.

12. Ziller, M.J., et al., *Charting a dynamic DNA methylation landscape of the human genome.* Nature, 2013. **500**(7463): p. 477-81.

13. Bird, A.P., *CpG-rich islands and the function of DNA methylation.* Nature, 1986. **321**(6067): p. 209-13.

14. Meissner, A., et al., *Genome-scale DNA methylation maps of pluripotent and differentiated cells.* Nature, 2008. **454**(7205): p. 766-70.

15. Watt, F. and P.L. Molloy, *Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter.* Genes Dev, 1988. **2**(9): p. 1136-43.

16. Bock, C., et al., *DNA methylation dynamics during in vivo differentiation of blood and skin stem cells.* Mol Cell, 2012. **47**(4): p. 633-47.

17. Xu, G.L., et al., *Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene.* Nature, 1999. **402**(6758): p. 187-91.

18. Gruenbaum, Y., H. Cedar, and A. Razin, *Substrate and sequence specificity of a eukaryotic DNA methylase.* Nature, 1982. **295**(5850): p. 620-2.

19. Okano, M., S. Xie, and E. Li, *Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases.* Nat Genet, 1998. **19**(3): p. 219-20.

20. Okano, M., Bell, D. W., Haber, D. A., Li, E., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-57.

21. Chen, T., N. Tsujimoto, and E. Li, *The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin.* Mol Cell Biol, 2004. **24**(20): p. 9048-58.

22. Sheikh, M.A., et al., *Epigenetic regulation of Dpp6 expression by Dnmt3b and its novel role in the inhibition of RA induced neuronal differentiation of P19 cells.* PLoS One, 2013. **8**(2): p. e55826.

23. Liang, G., et al., *Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements.* Mol Cell Biol, 2002. **22**(2): p. 480-91.

24. Bartolomei, M.S. and S.M. Tilghman, *Genomic imprinting in mammals.* Annu Rev Genet, 1997. **31**: p. 493-525.

25. Smith, Z.D. and A. Meissner, *DNA methylation: roles in mammalian development.* Nat Rev Genet, 2013. **14**(3): p. 204-20.

26. Lienert, F., et al., *Identification of genetic elements that autonomously determine DNA methylation states.* Nat Genet, 2011. **43**(11): p. 1091-7.

27. Ooi, S.K. and T.H. Bestor, *The colorful history of active DNA demethylation.* Cell, 2008. **133**(7): p. 1145-8.

28. Ito, S., et al., *Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification.* Nature, 2010. **466**(7310): p. 1129-33.

29. Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.

30. Pastor, W.A., L. Aravind, and A. Rao, *TETonic shift: biological roles of TET proteins in DNA demethylation and transcription.* Nat Rev Mol Cell Biol, 2013. **14**(6): p. 341-56.

31. Chung, S.Y., W.E. Hill, and P. Doty, *Characterization of the histone core complex.* Proc Natl Acad Sci U S A, 1978. **75**(4): p. 1680-4.

32. Gardner, K.E., C.D. Allis, and B.D. Strahl, *Operating on chromatin, a colorful language where context matters.* J Mol Biol, 2011. **409**(1): p. 36-46.

33. Ujvari, A., et al., *Histone N-terminal tails interfere with nucleosome traversal by RNA polymerase II.* J Biol Chem, 2008. **283**(47): p. 32236-43.

34. Fletcher, T.M. and J.C. Hansen, *Core histone tail domains mediate oligonucleosome folding and nucleosomal DNA organization through distinct molecular mechanisms.* J Biol Chem, 1995. **270**(43): p. 25359-62.

35. Struhl, K., *Histone acetylation and transcriptional regulatory mechanisms.* Genes Dev, 1998. **12**(5): p. 599-606.

36. Riggs, M.G., et al., *n-Butyrate causes histone modification in HeLa and Friend erythroleukaemia cells.* Nature, 1977. **268**(5619): p. 462-4.

37. Sealy, L. and R. Chalkley, *The effect of sodium butyrate on histone modification.* Cell, 1978. **14**(1): p. 115-21.

38. Anderson, J.D., P.T. Lowary, and J. Widom, *Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites.* J Mol Biol, 2001. **307**(4): p. 977-85.

39. Schubeler, D., et al., *Nuclear localization and histone acetylation: a pathway for chromatin opening and transcriptional activation of the human beta-globin locus.* Genes Dev, 2000. **14**(8): p. 940-50.

40. Kuo, M.H., et al., *Transcription-linked acetylation by Gcn5p of histones H3 and H4 at specific lysines.* Nature, 1996. **383**(6597): p. 269-72.

41. Turner, B.M., A.J. Birley, and J. Lavender, *Histone H4 isoforms acetylated at specific lysine residues define individual chromosomes and chromatin domains in Drosophila polytene nuclei.* Cell, 1992. **69**(2): p. 375-84.

42. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.

43. Rada-Iglesias, A., et al., *A unique chromatin signature uncovers early developmental enhancers in humans.* Nature, 2011. **470**(7333): p. 279-83.

44. Bannister, A.J. and T. Kouzarides, *Regulation of chromatin by histone modifications.* Cell Res, 2011. **21**(3): p. 381-95.

45. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.

46. Margueron, R. and D. Reinberg, *The Polycomb complex PRC2 and its mark in life.* Nature, 2011. **469**(7330): p. 343-9.

47. Strunnikova, M., et al., *Chromatin inactivation precedes de novo DNA methylation during the progressive epigenetic silencing of the RASSF1A promoter.* Mol Cell Biol, 2005. **25**(10): p. 3923-33.

48. Lehnertz, B., et al., *Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin.* Curr Biol, 2003. **13**(14): p. 1192-200.

49. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.

50. Ringrose, L., H. Ehret, and R. Paro, *Distinct contributions of histone H3 lysine 9 and 27 methylation to locus-specific stability of polycomb complexes.* Mol Cell, 2004. **16**(4): p. 641-53.

51. Francis, N.J., R.E. Kingston, and C.L. Woodcock, *Chromatin compaction by a polycomb group protein complex.* Science, 2004. **306**(5701): p. 1574-7.

52. Bernstein, B.E., et al., *Methylation of histone H3 Lys 4 in coding regions of active genes.* Proc Natl Acad Sci U S A, 2002. **99**(13): p. 8695-700.

53. Liang, G., et al., *Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome.* Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7357-62.

54. Pokholok, D.K., et al., *Genome-wide map of nucleosome acetylation and methylation in yeast.* Cell, 2005. **122**(4): p. 517-27.

55. Pray-Grant, M.G., et al., *Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation.* Nature, 2005. **433**(7024): p. 434-8.

56. Santos-Rosa, H., et al., *Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin.* Mol Cell, 2003. **12**(5): p. 1325-32.

57. Sims, R.J., 3rd, Millhouse, S., Chen, C. F., Lewis, B. A., Erdjument-Bromage, H., Tempst, P., Manley, J. L., Reinberg, D., *Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing.* Mol Cell, 2007. **28**(4): p. 665-76.

58. Breen, T.R., *Mutant alleles of the Drosophila trithorax gene produce common and unusual homeotic and other developmental phenotypes.* Genetics, 1999. **152**(1): p. 319-44.

59. Lee, J., et al., *Targeted inactivation of MLL3 histone H3-Lys-4 methyltransferase activity in the mouse reveals vital roles for MLL3 in adipogenesis.* Proc Natl Acad Sci U S A, 2008. **105**(49): p. 19229-34.

60. Jiang, H., et al., *Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains.* Cell, 2011. **144**(4): p. 513-25.

61. Dou, Y., et al., *Regulation of MLL1 H3K4 methyltransferase activity by its core components.* Nat Struct Mol Biol, 2006. **13**(8): p. 713-9.

62. Steger, D.J., et al., *DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells.* Mol Cell Biol, 2008. **28**(8): p. 2825-39.

63. Mohn, F., et al., *Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors.* Mol Cell, 2008. **30**(6): p. 755-66.

64. Li, B.Z., et al., *Histone tails regulate DNA methylation by allosterically activating de novo methyltransferase.* Cell Res, 2011. **21**(8): p. 1172-81.

65. Fuks, F., et al., *The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase.* Nucleic Acids Res, 2003. **31**(9): p. 2305-12.

66. Epsztejn-Litman, S., et al., *De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes.* Nat Struct Mol Biol, 2008. **15**(11): p. 1176-83.

67. Feldman, N., et al., *G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis.* Nat Cell Biol, 2006. **8**(2): p. 188-94.

68. Hahn, M.A., et al., *Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks.* PLoS One, 2011. **6**(4): p. e18844.

69. Azuara, V., et al., *Chromatin signatures of pluripotent cell lines.* Nat Cell Biol, 2006. **8**(5): p. 532-8.

70. Bernstein, B.E., et al., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells.* Cell, 2006. **125**(2): p. 315-26.

71. Voigt, P., W.W. Tee, and D. Reinberg, *A double take on bivalent promoters.* Genes Dev, 2013. **27**(12): p. 1318-38.

72. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state.* Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.

73. Voigt, P., et al., *Asymmetrically modified nucleosomes.* Cell, 2012. **151**(1): p. 181-93.

74. Schmitges, F.W., et al., *Histone methylation by PRC2 is inhibited by active chromatin marks.* Mol Cell, 2011. **42**(3): p. 330-41.

75. Ramsahoye, B.H., et al., *Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a.* Proc Natl Acad Sci U S A, 2000. **97**(10): p. 5237-42.

76. Ziller, M.J., et al., *Genomic distribution and inter-sample variation of non-CpG methylation across human cell types.* PLoS Genet, 2011. **7**(12): p. e1002389.

77. Xie, W., et al., *Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome.* Cell, 2012. **148**(4): p. 816-31.

78. Lister, R., et al., *Global epigenomic reconfiguration during mammalian brain development.* Science, 2013. **341**(6146): p. 1237905.

79. Clouaire, T., et al., *Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells.* Genes Dev, 2012. **26**(15): p. 1714-28.

80. Deng, C., et al., *USF1 and hSET1A mediated epigenetic modifications regulate lineage differentiation and HoxB4 transcription.* PLoS Genet, 2013. **9**(6): p. e1003524.

81. Ringrose, L. and R. Paro, *Polycomb/Trithorax response elements and epigenetic memory of cell identity.* Development, 2007. **134**(2): p. 223-32.

82. Mendenhall, E.M., et al., *GC-rich sequence elements recruit PRC2 in mammalian ES cells.* PLoS Genet, 2010. **6**(12): p. e1001244.

83. Arnold, P., et al., *Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting.* Genome Res, 2013. **23**(1): p. 60-73.

84. Vincent, S.D., et al., *Cell fate decisions within the mouse organizer are governed by graded Nodal signals.* Genes Dev, 2003. **17**(13): p. 1646-62.

85. Massague, J., *TGFbeta signalling in context.* Nat Rev Mol Cell Biol, 2012. **13**(10): p. 616-30.

86. Varelas, X., et al., *The Crumbs complex couples cell density sensing to Hippo-dependent control of the TGF-beta-SMAD pathway.* Dev Cell, 2010. **19**(6): p. 831-44.

87. Ishida, W., et al., *Smad6 is a Smad1/5-induced smad inhibitor. Characterization of bone morphogenetic protein-responsive element in the mouse Smad6 promoter.* J Biol Chem, 2000. **275**(9): p. 6075-9.

88. Dennler, S., et al., *Direct binding of Smad3 and Smad4 to critical TGF beta-inducible elements in the promoter of human plasminogen activator inhibitor-type 1 gene.* EMBO J, 1998. **17**(11): p. 3091-100.

89. Levy, L. and C.S. Hill, *Smad4 dependency defines two classes of transforming growth factor {beta} (TGF-{beta}) target genes and distinguishes TGF-{beta}-induced epithelial-mesenchymal transition from its antiproliferative and migratory responses.* Mol Cell Biol, 2005. **25**(18): p. 8108-25.

90. Germain, S., et al., *Homeodomain and winged-helix transcription factors recruit activated Smads to distinct promoter elements via a common Smad interaction motif.* Genes Dev, 2000. **14**(4): p. 435-51.

91. Seoane, J., et al., *Integration of Smad and forkhead pathways in the control of neuroepithelial and glioblastoma cell proliferation.* Cell, 2004. **117**(2): p. 211-23.

92. Ross, S., et al., *Smads orchestrate specific histone modifications and chromatin remodeling to activate transcription.* EMBO J, 2006. **25**(19): p. 4490-502.

93. Kang, Y., C.R. Chen, and J. Massague, *A self-enabling TGFbeta response coupled to stress signaling: Smad engages stress response factor ATF3 for Id1 repression in epithelial cells.* Mol Cell, 2003. **11**(4): p. 915-26.

94. Trompouki, E., et al., *Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration.* Cell, 2011. **147**(3): p. 577-89.

95. Mullen, A.C., et al., *Master transcription factors determine cell-type-specific responses to TGF-beta signaling.* Cell, 2011. **147**(3): p. 565-76.

96. Bilic, J., et al., *Wnt induces LRP6 signalosomes and promotes dishevelled-dependent LRP6 phosphorylation.* Science, 2007. **316**(5831): p. 1619-22.

97. Sierra, J., et al., *The APC tumor suppressor counteracts beta-catenin activation and H3K4 methylation at Wnt target genes.* Genes Dev, 2006. **20**(5): p. 586-600.

98. Graf, T. and T. Enver, *Forcing cells to change lineages.* Nature, 2009. **462**(7273): p. 587-94.

99. Gurdon, J.B., T.R. Elsdale, and M. Fischberg, *Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei.* Nature, 1958. **182**(4627): p. 64-5.

100. Taberlay, P.C., Kelly, T. K., Liu, C. C., You, J. S., De Carvalho, D. D., Miranda, T. B., Zhou, X., J., Liang, G., Jones, P. A., *Polycomb-repressed genes have permissive enhancers that initiate reprogramming.* Cell, 2011. **147**(6): p. 1283-94.

101. Struhl, K. and E. Segal, *Determinants of nucleosome positioning.* Nat Struct Mol Biol, 2013. **20**(3): p. 267-73.

102. You, J.S., Kelly, T. K., De Carvalho, D. D., Taberlay, P. C., Liang, G., Jones, P. A., *OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes.* Proc Natl Acad Sci U S A, 2011. **108**(35): p. 14497-502.

103. Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W. W., van Schaik, F. M., Varier, R. A., Baltissen, M. P., Stunnenberg, H. G., Mann, M., Timmers, H. T., *Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4.* Cell, 2007. **131**(1): p. 58-69.

104. Rahl, P.B., Lin, C. Y., Seila, A. C., Flynn, R. A., McCuine, S., Burge, C. B., Sharp, P. A., Young, R. A., *c-Myc regulates transcriptional pause release.* Cell, 2010. **141**(3): p. 432-45.

105. Hatta, M. and L.A. Cirillo, *Chromatin opening and stable perturbation of core histone:DNA contacts by FoxO1.* J Biol Chem, 2007. **282**(49): p. 35583-93.

106. Cuesta, I., K.S. Zaret, and P. Santisteban, *The forkhead factor FoxE1 binds to the thyroperoxidase promoter during thyroid cell differentiation and modifies compacted chromatin structure.* Mol Cell Biol, 2007. **27**(20): p. 7302-14.

107. Cirillo, L.A., Lin, F. R., Cuesta, I., Friedman, D., Jarnik, M., Zaret, K. S., *Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4.* Mol Cell, 2002. **9**(2): p. 279-89.

108. Zaret, K.S. and J.S. Carroll, *Pioneer transcription factors: establishing competence for gene expression.* Genes Dev, 2011. **25**(21): p. 2227-41.

109. Crossley, M., M. Merika, and S.H. Orkin, *Self-association of the erythroid transcription factor GATA-1 mediated by its zinc finger domains.* Mol Cell Biol, 1995. **15**(5): p. 2448-56.

110. Ko, L.J. and J.D. Engel, *DNA-binding specificities of the GATA transcription factor family.* Mol Cell Biol, 1993. **13**(7): p. 4011-22.

111. Merika, M. and S.H. Orkin, *DNA-binding specificity of GATA family transcription factors.* Mol Cell Biol, 1993. **13**(7): p. 3999-4010.

112. Ambrosetti, D.C., C. Basilico, and L. Dailey, *Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites.* Mol Cell Biol, 1997. **17**(11): p. 6321-9.

113. Remenyi, A., Lins, K., Nissen, L. J., Reinbold, R., Scholer, H. R., Wilmanns, M., *Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers.* Genes Dev, 2003. **17**(16): p. 2048-59.

114. Jauch, R., Aksoy, I., Hutchins, A. P., Ng, C. K., Tian, X. F., Chen, J., Palasingam, P., Robson, P., Stanton, L. W., Kolatkar, P. R., *Conversion of Sox17 into a pluripotency*

*reprogramming factor by reengineering its association with Oct4 on DNA.* Stem Cells, 2011. **29**(6): p. 940-51.

115. Nowling, T.K., Johnson, L. R., Wiebe, M. S., Rizzino, A., *Identification of the transactivation domain of the transcription factor Sox-2 and an associated co-activator.* J Biol Chem, 2000. **275**(6): p. 3810-8.

116. Tsuda, M., Takahashi, S., Takahashi, Y., Asahara, H., *Transcriptional co-activators CREB-binding protein and p300 regulate chondrocyte-specific gene expression via association with Sox9.* J Biol Chem, 2003. **278**(29): p. 27224-9.

117. Aksoy, I., et al., *Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm.* EMBO J, 2013. **32**(7): p. 938-53.

118. Hirai, H., Tani, T., Katoku-Kikyo, N., Kellner, S., Karian, P., Firpo, M., Kikyo, N., *Radical acceleration of nuclear reprogramming by chromatin remodeling with the transactivation domain of MyoD.* Stem Cells, 2011. **29**(9): p. 1349-61.

119. Yan, J., Xu, L., Crawford, G., Wang, Z., Burgess, S. M., *The forkhead transcription factor FoxI1 remains bound to condensed mitotic chromosomes and stably remodels chromatin structure.* Mol Cell Biol, 2006. **26**(1): p. 155-68.

120. Roberts, S.B., N. Segil, and N. Heintz, *Differential phosphorylation of the transcription factor Oct1 during the cell cycle.* Science, 1991. **253**(5023): p. 1022-6.

121. Segil, N., S.B. Roberts, and N. Heintz, *Mitotic phosphorylation of the Oct-1 homeodomain and regulation of Oct-1 DNA binding activity.* Science, 1991. **254**(5039): p. 1814-6.

122. Caravaca, J.M., et al., *Bookmarking by specific and nonspecific binding of FoxA1 pioneer factor to mitotic chromosomes.* Genes Dev, 2013. **27**(3): p. 251-60.

123. Gifford, C.A., et al., *Transcriptional and epigenetic dynamics during specification of human embryonic stem cells.* Cell, 2013. **153**(5): p. 1149-63.

124. Chambers, S.M., et al., *Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling.* Nat Biotechnol, 2009. **27**(3): p. 275-80.

125. Atlasi, Y., et al., *Wnt signaling regulates the lineage differentiation potential of mouse embryonic stem cells through Tcf3 down-regulation.* PLoS Genet, 2013. **9**(5): p. e1003424.

126. Evseenko, D., et al., *Mapping the first stages of mesoderm commitment during differentiation of human embryonic stem cells.* Proc Natl Acad Sci U S A, 2010. **107**(31): p. 13742-7.

127. D'Amour, K.A., et al., *Efficient differentiation of human embryonic stem cells to definitive endoderm.* Nat Biotechnol, 2005. **23**(12): p. 1534-41.

128. Hay, D.C., et al., *Highly efficient differentiation of hESCs to functional hepatic endoderm requires ActivinA and Wnt3a signaling.* Proc Natl Acad Sci U S A, 2008. **105**(34): p. 12301-6.

129. Bock, C., et al., *Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines.* Cell, 2011. **144**(3): p. 439-52.

130. Teo, A.K., et al., *Pluripotency factors regulate definitive endoderm specification through eomesodermin.* Genes Dev, 2011. **25**(3): p. 238-50.

131. Chambers, S.M., et al., *Combined small-molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors.* Nat Biotechnol, 2012. **30**(7): p. 715-20.

132. DeLaForest, A., et al., *HNF4A is essential for specification of hepatic progenitors from human pluripotent stem cells.* Development, 2011. **138**(19): p. 4143-53.

133. Ericson, J., et al., *Pax6 controls progenitor cell identity and neuronal fate in response to graded Shh signaling.* Cell, 1997. **90**(1): p. 169-80.

134. Sander, M., et al., *Genetic analysis reveals that PAX6 is required for normal transcription of pancreatic hormone genes and islet development.* Genes Dev, 1997. **11**(13): p. 1662-73.

135. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nat Biotechnol, 2013. **31**(1): p. 46-53.

136. Huntriss, J., et al., *Expression of mRNAs for DNA methyltransferases and methyl-CpG-binding proteins in the human female germ line, preimplantation embryos, and embryonic stem cells.* Mol Reprod Dev, 2004. **67**(3): p. 323-36.

137. Ostler, K.R., et al., *Cancer cells express aberrant DNMT3B transcripts encoding truncated proteins.* Oncogene, 2007. **26**(38): p. 5553-63.

138. Robertson, K.D., et al., *The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors.* Nucleic Acids Res, 1999. **27**(11): p. 2291-8.

139. Revil, T., et al., *Alternative splicing is frequent during early embryonic development in mouse.* BMC Genomics, 2010. **11**: p. 399.

140. Gordon, C.A., S.R. Hartono, and F. Chedin, *Inactive DNMT3B Splice Variants Modulate De Novo DNA Methylation.* PLoS One, 2013. **8**(7): p. e69486.

141. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.

142. Xie, W., et al., *Epigenomic analysis of multilineage differentiation of human embryonic stem cells.* Cell, 2013. **153**(5): p. 1134-48.

143. Ulitsky, I., et al., *Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.* Cell, 2011. **147**(7): p. 1537-50.

144. Mikkelsen, T.S., et al., *Comparative epigenomic analysis of murine and human adipogenesis.* Cell, 2010. **143**(1): p. 156-69.

145. Ahlgren, U., et al., *Independent requirement for ISL1 in formation of pancreatic mesenchyme and islet cells.* Nature, 1997. **385**(6613): p. 257-60.

146. Pfaff, S.L., et al., *Requirement for LIM homeobox gene Isl1 in motor neuron generation reveals a motor neuron-dependent step in interneuron differentiation.* Cell, 1996. **84**(2): p. 309-20.

147. Cai, C.L., et al., *Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart.* Dev Cell, 2003. **5**(6): p. 877-89.

148. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.

149. Spilianakis, C.G., et al., *Interchromosomal associations between alternatively expressed loci.* Nature, 2005. **435**(7042): p. 637-45.

150. Yeom, Y.I., et al., *Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells.* Development, 1996. **122**(3): p. 881-94.

151. Carey, B.W., et al., *Reprogramming factor stoichiometry influences the epigenetic state and biological properties of induced pluripotent stem cells.* Cell Stem Cell, 2011. **9**(6): p. 588-98.

152. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.

153. Lauberth, S.M., et al., *H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation.* Cell, 2013. **152**(5): p. 1021-1036.

154. Kwak, H., et al., *Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing.* Science, 2013. **339**(6122): p. 950-953.

155. Bock, C., et al., *Quantitative comparison of genome-wide DNA methylation mapping technologies.* Nat Biotechnol, 2010. **28**(10): p. 1106-14.

156. Norris, D.P. and E.J. Robertson, *Asymmetric and node-specific nodal expression patterns are controlled by two distinct cis-acting regulatory elements.* Genes Dev, 1999. **13**(12): p. 1575-88.

157. Saijoh, Y., et al., *Left-right asymmetric expression of lefty2 and nodal is induced by a signaling pathway that includes the transcription factor FAST2.* Mol Cell, 2000. **5**(1): p. 35-47.

158. Norris, D.P., et al., *The Foxh1-dependent autoregulatory enhancer controls the level of Nodal signals in the mouse embryo.* Development, 2002. **129**(14): p. 3455-68.

159. Boyer, L.A., et al., *Core transcriptional regulatory circuitry in human embryonic stem cells.* Cell, 2005. **122**(6): p. 947-56.

160. Thomson, M., et al., *Pluripotency factors in embryonic stem cells regulate differentiation into germ layers.* Cell, 2011. **145**(6): p. 875-89.

161. Wang, Z., et al., *Distinct lineage specification roles for NANOG, OCT4, and SOX2 in human embryonic stem cells.* Cell Stem Cell, 2012. **10**(4): p. 440-54.

162. Kist, R., E. Greally, and H. Peters, *Derivation of a mouse model for conditional inactivation of Pax9.* Genesis, 2007. **45**(7): p. 460-4.

163. Kagey, M.H., et al., *Mediator and cohesin connect gene expression and chromatin architecture.* Nature, 2010. **467**(7314): p. 430-5.

164. Yu, M., et al., *Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome.* Cell, 2012. **149**(6): p. 1368-80.

165. Spruijt, C.G., et al., *Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives.* Cell, 2013. **152**(5): p. 1146-59.

166. Xu, J., et al., *Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells.* Proc Natl Acad Sci U S A, 2007. **104**(30): p. 12377-82.

167. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nat Genet, 2007. **39**(4): p. 457-66.

168. Wamstad, J.A., et al., *Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage.* Cell, 2012. **151**(1): p. 206-20.

169. Koche, R.P., et al., *Reprogramming factor expression initiates widespread targeted chromatin remodeling.* Cell Stem Cell, 2011. **8**(1): p. 96-105.

170. Wang, J., et al., *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.* Genome Res, 2012. **22**(9): p. 1798-812.

171. Levinson-Dushnik, M. and N. Benvenisty, *Involvement of hepatocyte nuclear factor 3 in endoderm differentiation of embryonic stem cells.* Mol Cell Biol, 1997. **17**(7): p. 3817-22.

172. Dufort, D., et al., *The transcription factor HNF3beta is required in visceral endoderm for normal primitive streak morphogenesis.* Development, 1998. **125**(16): p. 3015-25.

173. Cirillo, L.A. and K.S. Zaret, *An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA.* Mol Cell, 1999. **4**(6): p. 961-9.

174. Brinkman, A.B., et al., *Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk.* Genome Res, 2012. **22**(6): p. 1128-38.

175. Koche, R.P., Smith, Z. D., Adli, M., Gu, H., Ku, M., Gnirke, A., Bernstein, B. E., Meissner, A., *Reprogramming factor expression initiates widespread targeted chromatin remodeling.* Cell Stem Cell, 2011. **8**(1): p. 96-105.

176. Shim, E.Y., C. Woodcock, and K.S. Zaret, *Nucleosome positioning by the winged helix transcription factor HNF3.* Genes Dev, 1998. **12**(1): p. 5-10.

177. Cirillo, L.A., et al., *Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4.* Mol Cell, 2002. **9**(2): p. 279-89.

178. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

179. Que, J., et al., *Multiple dose-dependent roles for Sox2 in the patterning and differentiation of anterior foregut endoderm.* Development, 2007. **134**(13): p. 2521-31.

180. Arnold, K., et al., *Sox2(+) adult stem and progenitor cells are important for tissue regeneration and survival of mice.* Cell Stem Cell, 2011. **9**(4): p. 317-29.

181. Shalek, A.K., et al., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.* Nature, 2013. **498**(7453): p. 236-40.

182. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

183. Ganga, M., et al., *PITX2 isoform-specific regulation of atrial natriuretic factor expression: synergism and repression with Nkx2.5.* J Biol Chem, 2003. **278**(25): p. 22437-45.

184. Tilgner, H., et al., *Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs.* Genome Res, 2012. **22**(9): p. 1616-25.

185. Wang, D., et al., *Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA.* Nature, 2011. **474**(7351): p. 390-4.

186. Kriks, S., et al., *Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson's disease.* Nature, 2011. **480**(7378): p. 547-51.

187. Whyte, W.A., et al., *Master transcription factors and mediator establish super-enhancers at key cell identity genes.* Cell, 2013. **153**(2): p. 307-19.

188. van Arensbergen, J., et al., *Derepression of Polycomb targets during pancreatic organogenesis allows insulin-producing beta-cells to adopt a neural gene activity program.* Genome Res, 2010. **20**(6): p. 722-32.

189. Nock, A., et al., *Identification of DNA-dependent protein kinase as a cofactor for the forkhead transcription factor FoxA2.* J Biol Chem, 2009. **284**(30): p. 19915-26.

190. Hynes, N.E., et al., *Signalling change: signal transduction through the decades.* Nat Rev Mol Cell Biol, 2013. **14**(6): p. 393-8.

191. Rahl, P.B., et al., *c-Myc regulates transcriptional pause release.* Cell, 2010. **141**(3): p. 432-45.

192. Huang, Y., et al., *The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing.* PLoS One, 2010. **5**(1): p. e8888.

193. Szulwach, K.E., et al., *Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells.* PLoS Genet, 2011. **7**(6): p. e1002154.

194. Costa, Y., et al., *NANOG-dependent function of TET1 and TET2 in establishment of pluripotency.* Nature, 2013. **495**(7441): p. 370-4.

195. Wu, H., et al., *Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells.* Nature, 2011. **473**(7347): p. 389-93.

196. Lagha, M., et al., *Paused Pol II coordinates tissue morphogenesis in the Drosophila embryo.* Cell, 2013. **153**(5): p. 976-87.

197. Apostolou, E., et al., *Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming.* Cell Stem Cell, 2013. **12**(6): p. 699-712.

198. Phillips-Cremins, J.E., et al., *Architectural protein subclasses shape 3D organization of genomes during lineage commitment.* Cell, 2013. **153**(6): p. 1281-95.

199. Konermann, S., et al., *Optical control of mammalian endogenous transcription and epigenetic states.* Nature, 2013. **500**(7463): p. 472-6.

200. Melnikov, A., et al., *Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.* Nat Biotechnol, 2012. **30**(3): p. 271-7.

201. Neph, S., et al., *An expansive human regulatory lexicon encoded in transcription factor footprints.* Nature, 2012. **489**(7414): p. 83-90.

202. Guelen, L., et al., *Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.* Nature, 2008. **453**(7197): p. 948-51.

203. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.

204. Weisenberger, D.J., et al., *Role of the DNA methyltransferase variant DNMT3b3 in DNA methylation.* Mol Cancer Res, 2004. **2**(1): p. 62-72.

205. Heyn, H., et al., *Whole-genome bisulfite DNA sequencing of a DNMT3B mutant patient.* Epigenetics, 2012. **7**(6): p. 542-50.

206. Espuny-Camacho, I., et al., *Pyramidal neurons derived from human pluripotent stem cells integrate efficiently into mouse brain circuits in vivo.* Neuron, 2013. **77**(3): p. 440-56.

207. Press, W.H., *Numerical recipes : the art of scientific computing*. 3rd ed2007, Cambridge, UK ; New York: Cambridge University Press. xxi, 1235 p.

208. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

209. Thorvaldsdottir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2012.

210. Zhu, L.J., et al., *ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data.* BMC Bioinformatics, 2010. **11**: p. 237.

211. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.

212. Kunarso, G., et al., *Transposable elements have rewired the core regulatory network of human embryonic stem cells.* Nat Genet, 2010. **42**(7): p. 631-4.

213. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif.* Bioinformatics, 2011. **27**(7): p. 1017-8.

214. Levin, J.Z., et al., *Comprehensive comparative analysis of strand-specific RNA sequencing methods.* Nat Methods, 2010. **7**(9): p. 709-15.

215. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.

216. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.